Edited by
Thomas Walker · Frederick Davis
Tyler Schwartz

Big Data in Finance
Opportunities and Challenges of Financial Digitalization

palgrave
macmillan

Big Data in Finance

Thomas Walker · Frederick Davis ·
Tyler Schwartz
Editors

Big Data in Finance

Opportunities and Challenges of Financial Digitalization

palgrave
macmillan

Editors
Thomas Walker
Department of Finance
Concordia University
Montreal, QC, Canada

Frederick Davis
Department of Finance
Concordia University
Montreal, QC, Canada

Tyler Schwartz
Department of Finance
Concordia University
Montreal, QC, Canada

ISBN 978-3-031-12239-2 ISBN 978-3-031-12240-8 (eBook)
https://doi.org/10.1007/978-3-031-12240-8

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Our global society is becoming increasingly data centric. The use of large, detailed datasets has revolutionized many fields, including—among others—medicine, biology, manufacturing, sports, marketing, and finance. With advances in how data can be collected and stored, a new phenomenon has emerged: big data. In simple terms, big data can be understood as a large amount of data that can be analyzed to understand past patterns better or predict future outcomes. This book examines the technical aspects of recent innovations surrounding big data in finance, as well as the benefits and risks associated with these developments. Moreover, the book sheds light on the ethical and privacy issues associated with big data, as well as the environmental footprint of collecting, storing, and analyzing big datasets.

The book features contributions from the international community of scholars and practitioners who work at the interface of artificial intelligence, big data, and finance. The authors review and critically analyze new developments at the intersection of big data and finance, and provide different perspectives on their impact on the financial sector and the way it operates. The book serves as a technical guide of these developments, exploring the theory and mechanisms behind the algorithms using big data, and exploring their use in a finance context. The contributors explain and demonstrate the predictive capabilities of big data in finance using different model types such as supervised, unsupervised, and semi-supervised learning. Moreover, because big data in finance has many applications that extend beyond financial institutions, the book features contributions that explore possible policy and sustainability-oriented solutions and implications of the use of big data in finance.

Montreal, Canada Thomas Walker
Montreal, Canada Frederick Davis
Montreal, Canada Tyler Schwartz

Acknowledgments

We acknowledge the financial support provided through the Jacques Ménard—BMO Centre for Capital Markets at Concordia University. In addition, we appreciate the excellent copy-editing and editorial assistance we received from Gabrielle Machnik-Kekesi, Victoria Kelly, Charlotte Frank, and Maya Michaeli.

Contents

Notes on Contributors

Abdool Imran is the President of consultancy Blue Krystal Technologies and Business Insights. Currently, he lectures at the University of Western Ontario Richard Ivey School of Business. During the 2007/2007 Financial Crisis, Imran also served in the Assistant Deputy Minister's Office for Financial Sector Policy of the Government of Canada's Department of Finance. Imran's commentaries on the economy and finance has appeared in Canada's national media such as the Globe and Mail, the CBC, and the Toronto Star.

Abdool Mustafa is a Machine Learning Engineer at Airbnb, one of the largest travel accommodations platforms in the world. His work involves designing and implementing novel recommender and search systems for Airbnb products. Academic papers authored by Mustafa and his team have appeared in the SIGKDD Conference on Knowledge Discovery and Data Mining (KDD). Mustafa holds a graduate degree in Computer Science, with a specialization in Artificial Intelligence from Stanford University.

Bhattacharya Prasanta is a research scientist and innovation lead with the Social and Cognitive Computing department at the A*STAR Institute of High Performance Computing. He also serves as adjunct Assistant Professor at the National University of Singapore (NUS) Business School. Prasanta holds a Ph.D. in Information Systems from the Department of Information Systems and Analytics, NUS, where he studied network science with a special focus on predictive and inferential methods in large social networks. His current research aims at understanding the role of big data in emerging social and business applications from finance and education. Prasanta actively collaborates with major industry partners from around the world, and has presented his research in leading computer science, information systems, and marketing science venues.

Bouabdallaoui Ibrahim is a Ph.D. candidate in Big Data and Artificial Intelligence at the Mohammadia School of Engineering, Mohammed V University in Rabat, Morocco. He holds an MSc in Big Data Analytics from Sidi Mohamed Ben Abdellah University in Fez, Morocco. His research interests include applications of artificial intelligence in tourism, finance, medicine and insurance.

Britez Isabel holds a degree in International Relations from the Federal University of Santa Catarina (UFSC), with graduate degrees in Neurosciences and Behaviour from the Pontifical Catholic University of Rio Grande do Sul (PUCRS) (Brazil). She currently pursues a Master of Studies in Social Innovation at the Judge Business School, University of Cambridge (class of 2024). She is Editorial Assistant for the Journal on Policy and Complex Systems, and Founder of ESG Means, a boutique research firm specialized in ESG investment.

Davis Frederick has been a professor at the John Molson School of Business since 2011, having received his Ph.D. shortly prior from Queen's University. Prior to his academic career, he worked for several years in the government sector focusing on economic development for communities. His research interests include mergers and acquisitions, insider trading, and other aspects of corporate finance. He has published in the Journal of Corporate Finance, Journal of Business Finance & Accounting, European Financial Management, the International Review of Financial Analysis, and other high-quality journals.

Fahmy Hany is Associate Professor and the Finance Intellectual Lead of the School of Business at Royal Roads University in Victoria, British Columbia. His research interests include energy economics, energy finance, financial econometrics, climate finance and economics, and sustainability. Fahmy's work has been published in top tier academic journals such as Energy Economics, Applied Economics, and Economic Modeling. His research work has been accepted and presented at top domestic and international economic and finance conferences. Fahmy's work has also been featured in the local media and other news outlets.

He is an active member of the Canadian Economic Association (CEA) and the Canadian Sustainable Finance Network (CSFN). In addition to his academic appointments, Fahmy has over ten years of experience in financial and economic consulting. He has led and authored a number of market studies, feasibility studies, cost-benefit analyses, and project appraisals for government and business.

Ferrouhi El Mehdi is Associate Professor of Finance at the Faculty of Economics and Management, Ibn Tofail University, Morocco. He holds a Ph.D. from the Mohammed V University in Rabat, Morocco. His research interests include behavioral finance, banking, risk management, financial intermediation, and financial inclusion. He has published papers in many prestigious journals including the Journal of African Business and Physica A.

Gobet Fernand is Professorial Research Fellow in the Centre for Philosophy of Natural & Social Science at the London School of Economics. His research interests include the psychology of expertise and talent, the acquisition of language, scientific discovery, and computational modeling. He has authored and co-authored over 400 publications, including eleven books. His latest books are *The Psychology of Chess* (2018) and *Scientific Discovery in the Social Sciences* (2019).

Grégoire Vincent is a tenured Associate Professor of Finance at HEC Montréal, where he holds the Research Professorship in Financial Big Data Analytics. He holds a Ph.D. in Finance from the University of British Columbia and degrees in Computer Engineering and Financial Engineering from Université Laval in Quebec. His academic research interests are in information economics, market microstructure, and big data and machine learning applications in finance. His research has been published in leading academic journals such as the Journal of Financial Economics, the Journal of Accounting Research, and the Journal of Financial and Quantitative Analysis.

Hameed Waseem Bak'r majored in Psychology and holds a Bachelor of Social Sciences (Honours) degree from the National University of Singapore. His varied background includes working in educational psychology and conducting neuroscience research with the Agency for Science, Technology, and Research (A*STAR). He has been Research Manager with Institute for Financial Literacy (IFL) since 2015.

Jepson Noah holds a B.B.A. in Finance and Economics from Bishop's University in Sherbrooke, Quebec where he was valedictorian of his graduating class. He is pursuing an MSc in Finance in Montreal. Noah's research interests are focused in the area of alternative data.

Kiat Quek Boon is a senior scientist and also director of the Social and Cognitive Computing Department at the A*STAR Institute of High Performance Computing. His research interests include computational modeling and simulation of psychological and cognitive processes, social network modeling and analyses, and building network-based modeling tools for knowledge representation and reasoning.

Peña Guillermo achieved recognition as the second highest graduating student of Economics for 2009–2013 at the University of Zaragoza (Unizar: Zaragoza, Spain), where he was also awarded the Extraordinary Prize for the highest grade in the Master of Economics program. He was a research visitor of the University of Oslo and obtained a Cum Laude International Ph.D. in Economics in 2018. Peña was employed in the Ibercaja Bank Headquarters for two years, and is currently Assistant Professor at the Department of Economic Analysis at the University of Zaragoza and researcher of the Institute for Employment, Digital Economy, and Sustainability (IEDIS). He has published more than 25 papers in prestigious journals such as the National Tax Journal and Applied Economics. His research fields include, but are not limited to, urban economics, public economics, and finance.

Purda Lynnette is Professor and Associate Dean (Graduate Programs) and RBC Fellow of Finance at the Smith School of Business, Queen's University, Kingston, Ontario. She is a Chartered Financial Analyst (CFA) with previous investment banking experience. Lynnette conducts empirical research that is frequently interdisciplinary in nature, with journal publications spanning the areas of accounting, finance, international business, and law. Outside of academia, Lynnette has presented her work to policy makers and practitioners including the Bank of Canada, the Accounting Standards Oversight Council of Canada, and Financial Planning Canada. Lynnette has been a visiting researcher at the Bank of Canada and INSEAD Business School (France) and organized workshops at the Halbert Centre for Canadian Studies at the Hebrew University of Jerusalem. Lynnette is a past president of the Northern Finance Association and an award-winning teacher.

Rathananthan Sivanithy has twenty-five years of experience as a financial journalist with The Business Times. For the majority of those years he was the newspaper's dedicated stock market reporter, writing daily and weekly market reports. He has written extensively about movements in the stock market and various financial instruments such as bonds, real estate investment trusts, and company-issued and structured warrants. He has served as a judge for the Securities Investors Association of Singapore's Corporate Governance awards. He holds a Bachelor of Accountancy degree from the National University of Singapore, a Master of Science (Management Science) from the University of Manchester, and a Master in Business Administration from the University of New South Wales.

Schwartz Tyler holds an MSc degree in Data Science and Business Analytics from HEC Montreal. He currently serves as a research assistant in the Department of Finance at Concordia University and is the co-author of an edited book collection on climate change adaptation. Tyler completed his undergraduate degree at the John Molson School of Business, for which he received an honors in finance, as well as a Concordia University Student Research Assistant (CUSRA) scholarship. His research interests include predictive modeling, bio-statistics, sports analytics, FinTech, and machine learning.

Seong Wan Kum is a Research Engineer for the Social and Cognitive Computing department at the A*STAR Institute of High Performance Computing. His research interests include personality and individual differences, behaviour and motivation, society and culture, as well as research methods & design.

Severino Federico is Assistant Professor in the Department of Finance, Insurance, and Real Estate at Université Laval, and a researcher at CIRANO. After obtaining his Bachelor and Master degrees in Applied Mathematics (with a focus on mathematical finance), he earned a Ph.D. in Economics and Finance from Università Bocconi (Milano) in 2018. He has also been a Postdoctoral Fellow at Università della Svizzera Italiana (Lugano). His research interests include asset pricing, financial economics, and financial econometrics. The main goal of his research is the study and classification of economic and financial shocks according to the duration of their effects on the markets. His research has been presented at several international conferences. Federico's teaching subjects include machine learning for business, portfolio management, corporate finance, asset pricing, and mathematics for economics.

Thierry Sébastien is a senior analyst at a major Canadian bank. He holds two Master's degrees: an M.B.A. in Finance from Université Laval and a Master of Science in Management from Grenoble École de Management. His expertise includes investments, econometrics, and corporate finance.

Tut Daniel holds a B.A. (Hons) in Economics from the University of Toronto and an M.A. in Economics and a Ph.D. in Finance from the Schulich School of Business at York University. Dr. Tut joined the Ted Rogers School of Business at Ryerson University in 2020. His research interests are in corporate finance, creditor rights, FinTech and cryptocurrencies, debt financing, and corporate governance. His work has been presented at some of the top finance conferences and he won the "Top Paper Award" at the 2021 Global Finance Conference. His work has been published in the Journal of Financial Research.

Venegas Percy is Chief Scientist at Economy Monitor, Editor-in-Chief at the Journal on Policy and Complex Systems (Policy Studies Organization, Washington, DC), and co-leader of Autonomous Hypotheses Generation for Investment Decision Making at the London School of Economics. He holds degrees in engineering, business, sustainability and a certification in Artificial Intelligence from Oxford University.

Walker Thomas is a full professor of finance, the director and academic lead of the Emerging Risks Information Center (ERIC), the inaugural director for the Jacques Ménard/BMO Center for Capital Markets, and the Concordia University Research Chair in Emerging Risk Management (Tier 1) at Concordia University in Montreal, Canada. He previously served as an associate dean, department chair, and director of Concordia's David O'Brien Centre for Sustainable Enterprise. Prior to his academic career, he worked for firms such as Mercedes Benz, KPMG, and Utility Consultants International. He has published over 70 journal articles and books.

Xu Zehuang is a graduate student of Business Analytics in the School of Business at Macau University of Science and Technology, Macau, where she also obtained her Bachelor of Applied Economics. His current research focuses on the application of machine learning methods in economics and finance.

Yang Wujin is a postgraduate student in Economics at the Hong Kong University of Science and Technology. Her undergraduate research in Applied Economics at Macau University of Science and Technology has given her a deep appreciation for advanced macroeconomics, advanced microeconomics, game theory, ESG investment, monetary policy, and information economics.

Yi Peng is a postgraduate student in the Department of Political Economy of King's College, London. She graduated from the Macau University of Science and Technology with a major in Applied Economics. During her undergraduate studies, her research involved combining economic theory with quantitative decision analysis methods to analyze the impact of the development of financial technology on the economies of different countries. During her postgraduate study, her research direction involves exploring the impact of fluctuations in the exchange rate of the RMB against the US dollar before and after the Sino-US trade war.

Ying Cecilia is a Ph.D. candidate in Analytics at the Smith School of Business at Queen's University. Her research explores the application of large language models in natural language processing tasks and their potential societal impacts. In addition, she is interested in machine learning and AI model explainability, biases, and corrections. Her work has been presented at leading computer science and management science conferences, including Neural Information Processing Systems, the Association for Computational Linguistics, and INFORMS. Cecilia is the recipient of the SSHRC Joseph-Armand Bombardier Canada Graduate Scholarship and the Smith School of Business New Ph.D. Student Research Excellence Award. Previously, Cecilia worked as a director of Credit Risk Analytics in Global Risk Management at Scotiabank.

Zhang John Fan currently serves as an Assistant Professor at Macau University of Science and Technology. John gained his Ph.D. in finance at Auckland University of Technology in 2018. His research interest lies in the area of multinational finance, the cross-listing of emerging market firms, organizational diversity, and the impact of culture on financial and economic aspects. He has published academic papers in the Review of Corporate Finance, International Research in Business and Finance, and China & World Economy. John is a Chartered Financial Analyst (CFA) and Microsoft Certified Solutions Expert (MCSE). John is also a member of the Christian Finance Faculty Association and treasurer of his local church.

Zhao Haorou is a Master of International Economics and Finance candidate at Newcastle University in the UK. She graduated from the Macau University of Science and Technology with a Bachelor of Applied Economics. In 2019, she participated in the Business Immersion program of The University of Hong Kong and the Global Business Case Competition. Her current research interests are in the area of macroeconomics, monetary policy, and forecasting.

List of Figures

Robo-Advisors: A Big Data Challenge

Bitcoin: Future or Fad?

Islamic Finance in Canada Powered by Big Data: A Case Study

Assessing the Carbon Footprint of Cryptoassets: Evidence from a Bivariate VAR Model

A Data-Informed Approach to Financial Literacy Enhancement Using Cognitive and Behavioral Analytics

List of Tables

High-Frequency Trading and Market Efficiency in the Moroccan Stock Market

Ensemble Models Using Symbolic Regression and Genetic Programming for Uncertainty Estimation in ESG and Alternative Investments

Robo-Advisors: A Big Data Challenge

Culture, Digital Assets, and the Economy: A Trans-National Perspective

Assessing the Carbon Footprint of Cryptoassets: Evidence from a Bivariate VAR Model

Introduction

Big Data in Finance: An Overview

Thomas Walker, Frederick Davis, and Tyler Schwartz

1 Introduction

In the financial sector, the big data movement refers to the analysis of vast amounts of data with the goal of making better informed investment decisions, improving corporate operations, and enhancing decision-making processes on both the buy and supply sides of transactions (Hasan et al., 2020). Big data analysis frequently draws on artificial intelligence (AI) models and has created a paradigm shift in the operations of financial institutions (Nobanee, 2021; Sun et al., 2019). Big data and its impact on the financial industry now benefit from scholarly interest; many publications, reports, books, and conventions focus on dissecting and comparing different applications and techniques within the domain,

T. Walker (✉) · F. Davis · T. Schwartz
Department of Finance, Concordia University, Montreal, QC, Canada
e-mail: twalker@jmsb.concordia.ca

F. Davis
e-mail: frederick.davis@concordia.ca

T. Schwartz
e-mail: tyler.schwartz@mail.concordia.ca

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_1

with researchers envisioning what the movement means for the future of finance (Nobanee, 2021). The two main subtopics of interest are big data itself as well as the AI techniques used to analyze that data.

AI is best understood as computer programming that automates tasks that have traditionally relied on human intelligence (Russell, 2010). A popular branch of AI is machine learning, which uses statistical techniques to allow machines to "learn" through experience by correcting previous errors (Alzubi et al., 2018). Spacecraft engineering, pattern recognition, entertainment, biology, finance, and medicine, among others, all rely on machine learning (El Naqa & Murphy, 2015).

More recently, deep learning, a sub-domain of machine learning, has proven to be effective in extracting patterns from big data (Chen & Lin, 2014). Deep learning uses neural networks, which are inspired by the neurological basis of brain functioning, to extract patterns from big data (LeCun et al., 2015). Primary applications of deep learning include image and voice recognition, as well as natural language processing (NLP) tasks such as text summarization and classification (LeCun et al., 2015; Najafabadi et al., 2015).

In the financial industry, AI has been successfully employed by financial institutions in such areas as automated lending, portfolio construction and management (robo-advising), risk management, fraud detection, quantitative and high-frequency trading, as well as customer support (Bahrammirzaee, 2010; Buchanan, 2019; Buchanan & Wright, 2021). As the complexity of algorithms increases with new developments in deep learning and other technological advancements, the role of AI in finance is certain to become more important as well, making the field a rapidly changing frontier in the financial technology (FinTech)/banking nexus and making space not only for opportunities, but also for substantial ethical, social, legal, and economic risks (Buchanan, 2019).

Big data, as a discipline, is what allows AI to learn and develop complex pattern-detecting algorithms used for a variety of purposes (O'Leary, 2013). The troves of information collected in finance, for example, include customer information on banking, historical prices of stocks for investing, social media activity to predict the current sentiment of the market, and past fraudulent transactions for fraud detection (Buchanan, 2019). The data can be structured (organized, classified data), unstructured (text, social media activity), or semi-structured (incorporating both structured and unstructured elements) (Hurwitz et al., 2013). However, most of today's data is unstructured and/or semi-structured, leading to greater difficulties in finding a way to use the data efficiently (Yaqoob et al., 2016).

Gandomi and Haider (2015) review the existing literature and define big data using the traditional three Vs (volume, variety, and velocity) along with three more recent Vs developed by International Business Machines (IBM) (veracity), Statistical Analysis System (SAS) (variability), and Oracle (value):

- **Volume**: the size of the data,
- **Variety**: the diversity of the data,
- **Velocity**: the speed at which the data is generated and processed,
- **Veracity**: the reliability of the data (for example, social media data),
- **Variability**: the variability in the velocity of data,
- **Value**: the quality and usefulness of the data.

Currently, many questions dominate big data scholarship and professional practice, including how data is collected (consumer privacy and ethical concerns), stored (environmental impacts), secured, as well as analyzed and used (Jain et al., 2016; Lucivero, 2020; Martin, 2015). This edited book explores new developments surrounding big data in finance, and aims to provide meaningful answers to the questions posed above. While a general use of big data has often been the subject of discussion, this book will take a more focused look at big data applications in the financial sector to differentiate itself from current offerings and close a large gap in the literature.

2 Overview of Content

This edited book critically analyzes new developments at the intersection of big data and finance and provides different perspectives on their impact on the financial sector and the way it operates. As illustrated above, the predictive capability of big data is still developing as increasingly complex algorithms are devised by institutions to extract the most out of available data. The chapters in this book explore the predictive ability of big data in the context of finance, using different model types such as supervised learning and unsupervised learning. In addition, several of the chapters discuss the leveraging of recent advancements made in deep learning in combination with big data to innovate the financial sector. Lastly, because the use of big data in finance has many implications that go beyond their use in financial institutions, several chapters address the ethical, privacy, and environmental implications of these applications.

2.1 Part I: Big Data in the Financial Markets

The book's first section looks at the use of big data in the financial markets, where it can be employed to optimize investment performance in various ways.

The section begins with Chapter 2, *Alternative Data*, where the authors discuss the role alternative data sources (e.g., internet activity, web traffic logs, chat boards, and social media platforms) can play in financial investment decisions. **Grégoire and Jepson** provide a critical analysis of the value gained from using such data sources against the difficulties in using that data—including the high cost associated with collecting, validating, and maintaining the data. The authors discuss the role big data can play in making alternative data more affordable to use.

Chapter 3, *An Algorithmic Trading Strategy to Balance Profitability and Risk,* proposes an algorithmic trading strategy (AT) based on financial indicators. The primary financial indicator this strategy is based on is the Balanced Investment Indicator (BII) which is used to find a balance between profitability and risk for an investor. **Peña** argues that using this proposed financial indicator for portfolio construction in combination with big data can provide an effective algorithmic trading strategy.

Following this, Chapter 4, *High-Frequency Trading and Market Efficiency in the Moroccan Stock Market,* explores whether high-frequency trading (HFT) impacts market efficiency. **Ferrouhi and Bouabdellaoui** employ a case study approach using data from the Moroccan Stock Exchange to evaluate this hypothesis at different time frequencies.

Lastly, Chapter 5, *Ensemble Models using Symbolic Regression and Genetic Programming for Uncertainty Estimation in ESG and Alternative Investments*, presents an ensemble algorithm using symbolic regressions (SR) to estimate the environmental, social, and governance (ESG) scores of publicly traded private equities (PE) and sustainable exchange-traded funds (ETFs). **Venegas, Britez, and Gobet** demonstrate how using the SR algorithm can reduce uncertainty in this investment space.

2.2 Part II: Big Data in Financial Services

The second section, *Big Data in Financial Services*, looks at different financial services that leverage big data, such as investing, credit, and digital assets, and how they are making an impact in the financial sector. In addition, this section presents how big data is, and can be, utilized in these financial services.

The section begins with Chapter 6, *Consumer Credit Assessments in the Age of Big Data*, which discusses how alternative data sources are being used in the financial services industry to produce credit assessments. **Purda and Ying** review and discuss the methodologies used to generate these credit quality assessments, which include machine learning methods. This chapter also presents privacy and ethical challenges linked to the use of these new alternative data sources when producing these credit assessments.

Chapter 7, *Robo-Advisors: A Big Data Challenge*, describes and evaluates how the financial services sector uses robo-advisors. **Severino and Thierry** discuss how this new financial innovation generates personalized portfolios for investors to optimize their financial returns. Lastly, this chapter discusses how big data, in conjunction with machine learning techniques, can improve the use and performance of these robo-advisors.

Chapter 8, *Bitcoin: Future or Fad?*, presents a holistic overview of Bitcoin's role in the financial market. **Tut** discusses critical questions concerning Bitcoin, which include its use as a cash proxy, its status as a store of value, its investment and diversification role, and its status as a collectible asset. Moreover, the chapter presents how Bitcoin, in conjunction with big data, can revolutionize industries such as healthcare and improve intellectual property rights. Lastly, the author discusses the role of government in regulating Bitcoin.

The concluding chapter in the section, Chapter 9, *Culture, Digital Assets, and the Economy: A Trans-National Perspective*, presents a statistical overview of the effect of culture on the use of digital assets. **Zhang, Xu, Peng, Yang, and Zhao** accomplish this by using an ordinary least squares (OLS) regression approach, processing data obtained from the World Bank database.

2.3 Part III: Case Studies and Applications

The book's final section, *Case Studies and Applications*, takes a case study approach to examine applications of big data in finance, and explores how these case studies serve as examples for the emerging global landscape. The section outlines how the case studies apply beyond their specific context and offers general recommendations regarding the use of big data in finance.

The first chapter in the section, Chapter 10, *Islamic Finance in Canada Powered by Big Data: A Case Study*, examines how big data and machine learning methods can, together, underpin the establishment of a credit union, the approvals for which are subject to extreme regulatory conditions. **I. Abdool and M. Abdool** present a case study of the Islamic credit union proposed in Toronto and how the union's founders can use big data tools to receive regulatory approval and capital from investors, and serve their market.

Following this, Chapter 11, *Assessing the Carbon Footprint of Cryptoassets: Evidence from a Bivariate VAR Model*, examines the carbon footprint of different cryptoassets, which include Bitcoin, Ethereum, Ripple, Stellar, and Litecoin. **Fahmy** uses a vector autoregression (VAR) approach, which models the relationship between the trading volumes of the cryptoassets mentioned above and their energy consumption. The chapter exposes one of the downsides of big data: its environmental footprint.

The book concludes with Chapter 12, *A Data-informed Approach to Financial Literacy Enhancement Using Cognitive & Behavioral Analytics*, in which the authors conduct a case study in Singapore to evaluate how big data, in combination with technological innovations, can improve the financial literacy of the population. **Bhattacharya, Seong, Kiat, Hameed, and Rathananthan** use the results of this case study to make recommendations to policymakers and educators around the globe on how to improve financial literacy programs to obtain better outcomes.

References

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine learning from theory to algorithms: An overview. *Journal of Physics: Conference Series, 1142*(1), 012012.

Bahrammirzaee, A. (2010). A comparative survey of artificial intelligence applications in Financial: Artificial neural networks, expert system and hybrid intelligent systems. *Neural Computing and Applications, 19*(8), 1165–1195.

Buchanan, B. G. (2019). *Artificial intelligence in finance*. The Alan Turing Institute.

Buchanan, B. G., & Wright, D. (2021). The impact of machine learning on UK financial services. *Oxford Review of Economic Policy, 37*(3), 537–563.

Chen, X.-W., & Lin, X. (2014). Big data deep learning: Challenges and perspectives. *IEEE Access, 2*, 514–525.

El Naqa, I., & Murphy, M. J. (2015). What is machine learning? In I. El Naqa, R. Li, & M. J. Murphy (Eds.), *Machine learning in radiation oncology* (pp. 3–11). Springer.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144.

Hasan, M., Popp, J., & Oláh, J. (2020). Current landscape and infuence of big data. *Journal of Big Data, 7*(1), 1–17.

Hurwitz, J. S., Nugent, A., Halper, F., & Kaufman, M. (2013). *Big data for dummies cheat sheet*. John Wiley & Sons.

Jain, P., Gyanchandani, M., & Khare, N. (2016). Big data privacy: A technological perspective and review. *Journal of Big Data, 3*(1), 1–25.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Lucivero, F. (2020). Big data, big waste? A reflection on the environmental sustainability of big data initiatives. *Science and Engineering Ethics, 26*(2), 1009–1030.

Martin, K. (2015). Ethical issues in the big data industry. *MIS Quarterly Executive*.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data, 2*(1), 1–21.

Nobanee, H. (2021). A bibliometric review of big data in finance. *Journal of Big Data, 9*(2), 73–78.

O'Leary, D. E. (2013). Artificial intelligence and big data. *IEEE Intelligent Systems, 28*(2), 96–99.

Russell, S. J. (2010). *Artificial intelligence: A modern approach*. Pearson Education, Inc.

Sun, Y., Shi, Y., & Zhang, Z. (2019). Finance big data: Management, analysis, and applications. *International Journal of Electronic Commerce, 23*(1), 9–11.

Yaqoob, I., Hashem, I. A., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management, 36*(6), 1231–1247.

Big Data in the Financial Markets

Alternative Data

Vincent Grégoire and Noah Jepson

1 Introduction

In finance, the use of big data involves analyzing vast amounts of data to make better-informed investment decisions, improve corporate operations, and enhance analysis on both the buy-side and sell-side. Traditionally structured data sources, such as financial trade and quote data, financial statements, earnings reports, sell-side transaction equity research, and economic data releases, are easy to use and widely available. While playing an essential role in any analysis, traditional financial data is so ubiquitous that gaining an informational edge is difficult. This lack of an advantage leads market participants to search for alternative sources of information to gain insight into investment decisions. Technology has made it easier and cheaper to collect alternative data in large quantities, which has led to an arms race in data acquisition.

V. Grégoire (✉) · N. Jepson
Department of Finance, HEC Montréal, Montréal, Canada
e-mail: vincent.3.gregoire@hec.ca

N. Jepson
e-mail: noah.jepson@hec.ca

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_2

This chapter begins by examining the rise of alternative financial data. Historically, producing such datasets was labor-intensive, for example, through the use of scouts to count foot traffic at retail venues or by taking photos of different job sites. However, with advances in big data technologies, new data sources can be collected, transferred, and processed with greater efficiency. Furthermore, a large stock of alternative datasets is now available for purchase from an ever-increasing pool of vendors.

Next, we describe the most common types of alternative financial data, their use cases, and limitations. A large segment of the alternative data landscape is occupied by data collected from internet activity. Advances in big data technologies have made it possible to easily collect and aggregate text data, images, and web traffic logs. The data may come from news articles, chat boards, e-commerce websites, or social media. A widespread use case for social media data is tracking the sentiment of market participants, derived using text analysis of user-generated content. Headcount and hiring can be monitored through online job listings, and online ratings can provide insight into a company's culture.

Finally, we discuss issues surrounding the evaluation of additional datasets. Because information provides a competitive advantage to financial actors, it is not easy to find detailed information regarding the value of an additional dataset. This first creates a reverse survivorship bias as to what firms will say publicly about alternative data sources because data sources that lead to the production of alpha or a risk reduction are less likely to be advertised by those using them. Second, a cost is attached to inaction because competitors invest in acquiring alternative financial data.

2 Characteristics of Alternative Data

Alternative financial data is commonly defined as any data used in financial decision-making but that are not traditional financial data, such as trade and quote data, financial statements, earnings reports, and analyst research. Denev and Amen (2020) give the following five main characteristics of alternative data. Table 1 details each of these characteristics and briefly summarizes each. We provide further analysis of each characteristic in the text which follows.

Table 1 Main characteristics of alternative data

Characteristic	*Reason*
Less commonly used by market participants	Alternative data gives a different perspective on investment decisions
Tend to be more costly to collect and purchase	Data must be sourced and stored
Typically outside of financial markets	In line with providing a different perspective, data usually stems from upstream of traditional data sources
Tend to lack historical data	Compared to traditional data, alternative dataset sources are younger
More challenging to use	Tend to be more unstructured; there is no clearcut way to use the data

2.1 Less Commonly Used by Market Participants

Alternative datasets are typically developed and acquired to provide an informational edge over competitors. For this edge to be valuable, the data should be unique or, at the very least, not in widespread use. However, one must accept the condition that if a dataset contains valuable information, this will inevitably lead to an increase in the popularity of that dataset. Therefore, once deemed “alternative,” a dataset can become a traditional financial dataset over time (Fleming-Williams, 2020).

2.2 Tend to Be More Costly to Collect and Purchase

Although alternative data is cheaper now than ever before, it remains very costly to collect, validate, and process. With the advent of big data, a considerable amount of data collection has become automated. However, the data collection and analysis process requires specialized labor, such as computer engineers and machine learning specialists. A growing industry of alternative data providers build, maintain, and distribute datasets to a large number of clients. This growth helps them to spread the development cost among multiple users. However, while this leads to reduced costs, it also leads to broader use of the data and thus reduces the competitive advantage associated with the use of the data.

2.3 Typically Outside of Financial Markets

Alternative financial data is typically sourced outside of financial markets. It can be derived from economic activity such as credit card transactions, human activity such as social network interactions, or other sources such as climate data.

2.4 Tend to Lack Historical Data

Because the popularity of alternative financial data is recent, the availability of historical data is quite limited. Furthermore, some data sources, such as social media or data from online transactions, have only existed for a short time.

2.5 More Challenging to Use

Most traditional financial data sources are of the structured kind. For example, trade and quote data and financial statements contain structured quantitative information that can be easily embedded into financial decision models. In contrast, many alternative data sources such as news articles, social media posts, and results from web scraping tend to be loosely structured or completely unstructured, making their integration into financial models challenging. Even structured alternative data, such as satellite imagery or weblogs, can be challenging to embed in financial models because the roles they play in those models are, as yet, unclear.

3 Catalysts of the Growth in Alternative Data

The first catalyst for the growth in alternative data is the digitalization trend. A whopping "2.5 quintillion bytes of data" are reportedly created each day (Marr, 2018). Some calculations indicate that 90% of all data in circulation has been created in the last two years, highlighting the dramatic growth of collectible data points. Cloud computing and the Internet of Things have opened up thousands of additional data points on people and their actions. As many processes are automated, a vast amount of "exhaust data" is produced (Liberti & Petersen, 2019). One example of exhaust data is the metadata that follows an online transaction, including the time of the transaction, location, and device used.

These exhaust data points can now be aggregated, tracked, and packaged as alternative data. These data collections tend to be extensive and contain various data types.

The second catalyst to the rise of alternative data is the decreasing cost of data storage. Why would anyone want to store this enormous amount of data, especially if it will not be converted into informed decisions or alpha returns? With the low cost and accessibility of data storage, the answer is often "why not?". Sharma (2018) outlines the decreasing cost of storing one gigabyte of data from a high of $200,000 in the 1980s, to $7.70 in the early 2000s, to just $0.02 in 2017. The ability to collect and archive data at such a low cost creates the "why not?" rationale when determining what and how much data to store.

Moreover, alternative data providers have reaped the benefits of lower-cost data storage and increased data points. There were an estimated 100 alternative data providers in 2008, compared to over 400 providers just ten years later in 2018 (*Industry Stats—AlternativeData.Org*, 2019). This increase is not surprising given the financial industry's exponential rise in spending on alternative data. Total spending from buy-side firms on alternative data increased from around $232 M in 2016 to a projected $1.708B in 2020.

Increased demand from financial markets has pushed the number of providers and spending up, but the question remains: who uses these data and why? The Financial Times has cited increased pressure from low-cost index funds as a primary driver for asset managers to turn to alternative data in their search for an edge (Wigglesworth, 2018). Hedge funds have historically been very active in the alternative data sector. A recent 2019 EY survey revealed that 80% of hedge funds with over $10 billion in assets under management (AUM) list alternative data vendors as their source of "next-generation data" (EY Global Alternative Fund Survey, 2019).

4 Sources of Alternative Data

Financial firms have embraced alternative data using two main approaches. The first is by assembling their own data analysis teams, which allows them to customize their approach to distilling massive amounts of data into valuable insights that inform decision-makers. The second is by purchasing alternative data providers' services to acquire alternative data that has been structured and can be easily added to financial models. In between these two approaches exists a wide array of possibilities, including purchasing unstructured data and processing it internally or collecting unique data and hiring third parties to help with the processing.

From collection to curation, alternative data can be costly. The 2017 EY Global Hedge Fund and Investor Survey highlights that high costs arise when investors choose to develop the ability to decipher relationships and signals from alternative data. This development is reflected in the strong demand for specialized labor, especially data engineers and data analysts. The prominent financial careers website eFinancialCareers reported a growth of 460% in the number of financial firms mentioning "alternative data" in job postings in the last six months of 2021 (Butcher, 2022). Maintaining these teams can prove to be a costly endeavor. AlternativeData.org estimated the cost of maintaining an in-house team to collect and analyze alternative data in 2018 to be in the $1.5 M to $2.5 M range (Tomaselli & Kilduff, 2018). The benefit of having an in-house data team is the ability to gather and analyze unique, unstructured data. The alternative is to buy an existing structured dataset from an alternative data provider. However, while third-party datasets are often cheaper, they tend to be less valuable because of their wider availability, lowering their competitive value.

Additionally, alternative data also come with risks. First, as alternative data becomes more critical to operations, it is essential to realize that data sources come and go, especially those that are not the by-product of regulations such as mandatory financial disclosures. Losing a data stream can cause downtime as the organization pivots to find a replacement dataset or recalibrate models to handle the missing variables. This risk is particularly acute with user-generated data because of the growing concerns surrounding privacy issues related to this data. Concerns regarding privacy practices represent a real threat to the alternative data industry as they can reduce data access for alternative data providers, users, and aggregators. In the November 2020 episode of "The Alternative Data Podcast," the Chief Revenue Officer of Factus, an alternative data provider, highlighted an event in which their data supplier threatened to pull back data sales, afraid of the possible backlash from users and the public (Mark Fleming-Williams, 2021). That put Factus in a difficult position with clients that relied on their datasets. Factus overcame this challenge by anonymizing the data, which removed all personally identifiable information.

Although most realizations of this risk are kept private, especially in the financial world, we observe the same risk in the tech and advertising industry. In most cases, the digital service provider owns the data, and the user must accept the terms of service agreement of the provider. These agreements typically allow the provider to use, collect, and sell the user data. However, regulators are taking notice, and new regulations that restrict the use of private data have been introduced or are planned. The most significant of these is Europe's General Data Protection Regulation (GDPR), which restricts how firms collect data and what they can do with them. Private firms are also catering to privacy-conscious customers with solutions such as tracking restrictions built into the leading web browsers, "no-data-collected" new search engines such as DuckDuckGo, and more robust privacy policies such as the one that Apple has introduced in their app store guidelines. Bloomberg.com reports that Facebook revenues are expected to drop by 7% due to Apple's recent iOS update that allows iPhone users to opt-out of app tracking; this drop is partly due to advertisers who previously relied on Facebook's ads to push products looking elsewhere in panic (Wagner, 2021).

These new government regulations and data privacy changes on Apple's iOS devices are just the beginning of data privacy reforms on both public and private levels. These changes are happening in part due to significant backlash from users regarding how their information is being used. Therefore, it is imperative not to take a data source for granted as company and government privacy laws and regulations constantly evolve.

5 Types of Alternative Data

Alternative data comes in many forms. The 2017 EY Global Hedge Fund and Investor Survey ranked the types of alternative data in current use: social media ranked highest, followed by credit card data in third place, with web traffic, search trends, point of sale data, and crowdsourced data all within the top 14 spots. This section describes some challenges associated with dealing with data in text form, which is typical for alternative data sources. We then examine the use cases, methods, and applications of some other common types of alternative data sources to shed light on their applications in the financial industry.

5.1 Text Data

Digitization has pushed many real-life processes into the digital sphere. As most information is collected in text form, the term text data is one of the more extensive umbrella terms for alternative data. One challenge when working with text data is the struggle to extract actionable information from unstructured or semi-structured data (Gentzkow et al., 2019); the less structured the data, the more costly it will be to extract valuable information from that data. Talib et al. (2016) estimated that 80% of data today are unstructured, so extracting information from the vast amount of text data is a complex challenge for end-users. This difficulty can influence the business decision to collect and process textual data, i.e., to invest in an in-house data team or purchase alternative datasets.

The classic approach for collecting textual data from the internet begins with identifying the target text data. Vetting the target website's terms and conditions and the company's overarching policies is imperative. Although text data can often be found on the surface of many websites, most platforms prohibit automated scraping; a breach of the website's terms and conditions may result in significant legal and reputational risk. Collecting the text data in a uniform and repeatable way is critical, regardless of whether the data are in PDF form, on a web page, or an image. In the next step, pre-processing and data cleaning are performed to detect and remove inconsistencies within the dataset (Gupta et al., 2020).

5.2 *Job Postings and Other Economic Activity Indicators*

Gutiérrez et al. (2020) consider how online job postings can provide insight into the company's growth. Human capital is an integral part of any modern corporation, and these researchers claim job postings are disclosers made outside the investor-relations channel that contain forward-looking information. They find "that changes in the number of job postings are positively associated with changes in future performance and that this relationship is stronger when postings likely represent growth rather than replacement" (Gutiérrez et al., 2020, p. 1). Employee count can be identified by growth versus replacement. These data suggest that access to job postings of specific companies can give a view of the future performance and growth of the company. Some related data sources include application ratings, company reviews, products by a vendor listed as a bestseller or discounted, and web searches. All of these data sources point to a similar story: the more data that can be collected from the public-facing website of the company or from a related third-party website, the more investors can gain a view of the reputation and selling behavior of the company, indicating how a company might perform in the future (Drake et al., 2012; Green et al., 2019).

5.3 *Mandatory Disclosures*

As disclosure requirements of firms and governmental entities increase, regulatory filings become an essential source of alternative data.

In October 2010, United States President Barrack Obama signed Senate Bill 3717 into law. This bill amended the "Investment Advisers Act of 1940 (the 'IAA') by repealing a broad exemption of the Securities and Exchange Commission ('SEC') from the Freedom of Information Act ('FOIA') and other disclosure obligations" (Bennett et al., 2010, p. 1). This bill opened the SEC for FOIA requests, which financial firms, notably hedge funds, began to use extensively. The SEC filing platform, Edgar, is bot-friendly, facilitating the data retrieval process. Furthermore, facing an increase in related FOIA requests, the SEC now makes publicly available the server logs in a semi-anonymized format, allowing everyone to see which documents were downloaded, when, and by whom based on the IP address range.

However, the FOIA ruling had implications reaching further than the SEC. In 2012, the leading government agencies that received FOIA requests were the Department of Homeland Security, the Department of Justice, and the Department of Health and Human Services (which oversees the Food and Drug Administration (FDA), the decision-maker in the approval of drugs and food recalls). Overall, requests have skyrocketed in the past decade, increasing by 80% from already high levels in 2012 to over 8 million requests in 2018 (Sager, 2020).

The Wall Street Journal (WSJ) has investigated large hedge funds' numerous requests and subsequent trades. In a 2013 article, the WSJ reported that SAC Capital Advisors and their parent affiliate Sigma Capital Management requested "any adverse reports" for biotech company Vertex's new cystic fibrosis drug (Mullins & Weaver, 2013). The request was granted under the FOIA, and documents providing evidence of "no major problems" were released to SAC. Subsequently, SAC purchased over 13,000 shares of Vertex and some options. Later that year, when the positive drug reports were finally publicly announced, the stock rose 62% in one day. This example shows the power of FOIA requests in gaining an additional perspective on potential investments.

Moreoverm, Gargano et al. (2017) explored this case further. The authors used FOIA requests to obtain a list of FOIA requests made by other institutions to the FDA. They found that FOIA requests were chiefly used by large, sophisticated investors such as hedge funds that "trade more frequently than their peers." The request usually surrounds "stocks that are complex to value, and in periods of high firm or market uncertainty" (Gargano et al., 2017, p. 47). Their study found that when FOIA requestors increased their holdings, the quarterly abnormal returns averaged 6.58%; when the requestor decreased holdings, the quarterly abnormal returns averaged −3.52%. Data aggregation and collection is a budding industry for the many reasons cited above, but privacy is an integral part of this subtopic of alternative text data. Connecting abnormal returns to specific firms, as reflected in lists of FOIA requests, effectively gives an investor the ability to snoop on another company's data practices and curiosity. Many firms (namely, in this study, FOI Services, Inc.) have started to file requests on behalf of investors to shield where a firm's interest may lie. The WSJ article notes that "FOI Services Inc. accounted for about 10% of the 50,000 information requests sent to the FDA during the period examined by the Journal" (p. 3), showing the sheer demand for FOI's services.

5.4 Social Media Data: Use Cases, Methods, Applications

Social media platforms such as Facebook, LinkedIn, Twitter, and Reddit generate many metrics on users and content such as follows, likes, comments, and other data points such as cookies, logs, and temporary internet files. Given the estimated 4.6 billion social media users worldwide, the importance of these data is obvious (Global Social Media Stats, 2022). However, not all user data are easy to collect or scrape. For example, Facebook has restricted access to its data following the Cambridge Analytica debacle. In contrast, Twitter facilitates the scraping of its platform for a fee.

Literature on the informational value of social media content has resulted in conflicting conclusions that have fluctuated over the years, as has the use of social media in everyday life. Tumarkin and Whitelaw (2001) investigated the link between finance-focused message boards and stock returns and volume. They found no causal link between message board activity and stock returns and volume, and concluded that the effect goes in the opposite direction: market information influences message board activity. Antweiler and Frank (2004) took to Yahoo! Finance and measured the "bullishness" of a particular message regarding a publicly-traded company using text analysis. They concluded that stock-related messages helped predict market volatility, even after controlling for other sources of information, such as the WSJ's news reports. Stock messages had a statistically significant effect on stock returns but were economically small.

Recent evidence suggests that social media is not only a reflection of investor sentiment or attention, but that it can be a driving force behind investment trends. In 2001, message boards and blogs were still growing in popularity, but their reach was quite limited. In early 2021, market events surrounding the "meme stock" craze triggered an investigation by the SEC into the dramatic increase in GameStop's stock price. The SEC investigation report, published in October 2021, touted the idea that GameStop, as well as other "meme stocks," experienced "price increases, trading interest, and social media interest [that] all accelerated in 2021" (SEC.Gov, 2021, p. 18). Moreover, in a 2022 survey on TikTok reported in The Economist, "nearly a quarter of investors aged 18 to 40, and 41% of those aged between 18 and 24 years old, have sought financial advice on the platform" ("Personal finance is a hit on TikTok," 2022, para. 1).

Duz Tan and Tas (2021) investigated the impact of social media, specifically on the S&P index constituents for the United States, Europe, and other emerging markets. By using firm-specific Twitter data and running tweets through sentiment analysis, they found that "firm-specific Twitter sentiment contains information for predicting stock returns, and this predictive power remains significant after controlling for news [report] sentiment" (p. 1). Their findings align with Blankespoor et al. (2013), who found that social media content produced by firms themselves has strong predictive power, which they attribute to investors being better informed after utilizing those additional disclosure channels.

Additionally, Cookson and Niessner (2020) explored the user-generated content of StockTwits, an investment-focused social media platform, to study the sources of investor disagreement online. For example, an investor might be more inclined to invest based on a technical or fundamental approach, which could cause disagreement in a stock's valuation. The authors separated disagreements stemming from separate information sets or different interpretations of information. They found that sources of disagreement on the platform were more closely related to trading volume when the disagreement was within investment approach groups, suggesting that information variations are more significant within than across investment approaches.

Most recent literature suggests that social media content can be a valuable predictor of volatility and trading volume.

5.5 Transaction Data

An increasingly large proportion of retail transactions are now settled electronically, generating valuable exhaust data. Payment providers use that data as another source of revenue, selling aggregated insights derived from credit cards and other electronic transactions. A 2018 Forbes article noted that Mastercard, AmEx, and Envestnet profit over $400 M annually from selling transaction data to private equity and investment banks (Cohan, 2018). These data are anonymized to preserve consumer privacy but still provide enough detail to draw out the aggregated trends or track consumers.

Furthermore, this kind of data can be used as a leading indicator of economic activity and consumer spending (Baker & Kueng, 2021). The data is now so detailed that they can also provide insights about company revenues and be used to determine consumer fidelity and predict lifetime customer value ("Predictive Modeling Using Transactional Data," 2010). These can be valuable insights for the future value of a company or firm. One of the reasons why credit card and transaction data are so sought after is their ability to be quickly updated and incorporated into the decision-making process. MapD, a data and analytics provider, provides the following note on their hedge fund clients: "[our clients] use MapD to analyze stock ticker data and credit card transactions—they analyze 10%–30% of all transactions in real-time to see which companies are enjoying upticks in purchases" (Cohan, 2018, para. 2). Many large banks invest in innovative alternative data companies, namely those specializing in transaction data (Goldman Sachs and Citi Make $20 Million Venture Bet That Private Equity Wants to Your Credit Card Info, 2019).

Although academic studies using this type of data are scarce, most likely due to exclusivity and cost, banks and credit card providers possess excellent reports on the impact of transaction data on economic activity. One such bank is Banco Bilbao Vizcaya Argentaria (BBVA), a financial services institution headquartered in Spain. BBVA conducted an internal application using transaction data to gain insight into the retail trade of Spanish citizens. In the report, BBVA investigated how transaction data could replace the Spanish Retail Volume Trade Index (RTI), a metric showing the evolution of aggregate consumption and economic activity. BBVA found that "card transaction data replicate with great precision the evolution of the Spanish RTI. "Furthermore, BBVA discovered that "working with an internal dataset allowed for great granularity leading to the ability to replicate the evolution of 'daily retail sales,' with timely answers on the impact of any retail sales event" (Bodas-Sagi et al., 2019, p. 31) The BBVA report was able to model retail trade with great geographical detail and even pin down the activity sector (Bodas-Sagi et al., 2019).

The quantity and granularity of transaction data are influential for many users. Venture capital funds, private equity firms, and investment banks use these datasets to gain revenue insights on private and public companies alike. Additionally, for more macro-focused hedge funds and financial service companies, transaction data give real-time insight into economic conditions for specific regions, sectors, or nations.

5.6 Satellite Imagery and Weather Data

Satellites provide a convenient way to monitor wide-scale human activity, reducing the effort required to collect information. For example, this type of data can be used to count the number of cars in a car park at any given store or location. Before the advent of commonly available low-cost satellite imagery, this type of information was costly to collect. For example, hedge funds would employ locals to monitor car parks in real-time. Satellite imagery can also be used to monitor all sorts of human and non-human activity, such as tracking trade activity by monitoring ports or tracking weather events in real-time.

A notable use of satellite imagery is in monitoring construction and infrastructure projects. Monk et al. (2018) outlined the use of images of shadow lengths from construction sites as a measure of building progress. Algorithms can convert the measurements into calculations of the pace of the project so that an investor might enjoy greater clarity on the building progress. This type of alternative data can reduce risk by giving concrete insights into potential construction delays.

Climate impacts not only a firm's activity but also human behavior. Ic, Kahyaoglu, and Odabas (2014) built on the documented relationship between mood, emotion, and feelings in decision-making using weather and temperature data. They explored abnormalities in the financial markets with a particular focus on non-financial variables, including investors' feelings and the weather. They found a negative relationship between cloudiness and the percentage of fixed-income securities held in a trader's portfolio (their measure of perceived risk level). Similarly, they found a negative relationship between average temperature and the percentage of purchases but a positive relationship between temperature and sales. These findings suggest that information to be found in climate data goes beyond firm performance and can also inform investors about the behavior of market participants (İç et al., 2014).

6 Processing Alternative Data

The growth in alternative data has been made possible by advances not only in data collection and decreasing storage costs, but also in machine learning technology, especially in the field of deep learning. These advances have allowed consumers of alternative data to automatically extract complex information that would have previously required interpretation by a human analyst.

In particular, innovations in natural language processing (NLP) methods have been instrumental to the growth of alternative data in finance. NLP is a typical application of artificial intelligence (AI) that aims to extract meaning from text data, such as those found on social media sites. Consider the typical use case of extracting sentiment or attention metrics from user-generated content expressing an interest in the economy or a specific firm on social media platforms. The signal would be derived from an interest in the company's products or directly from investment-oriented content, as investors increasingly rely on social media as an essential source of financial information. Using NLP allows the dataset to be mined to curate a set of insights and conclusions. In other words, NLP techniques enable computers to understand the text as if a human was reading it. These techniques were initially simple, with dictionary-based word count algorithms such as those popularized by Tetlock (2007) and Loughran and McDonald (2011). Over time, more sophisticated algorithms gained traction, such as the Latent Dirichlet Allocation method first introduced by Blei, Ng, and Jordan (2003), which years later became a popular method for extracting topics from financial text data. Recently, advances in transformer-based deep learning models derived from the BERT algorithm of Devlin et al. (2018) have shown incredible performance in extracting meaning from complex text and currently represent the state-of-the-art in the field.

Furthermore, NLP can be employed by users in many different domains, and the sheer quantity of text data that can be analyzed is one of NLP's key benefits. Financial actors can use NLP processes to draw insights from existing traditional sources of information, such as Managers' Discussion and Analysis documents or aggregated social media logs. The result is an efficient way to draw out latent conclusions and insights that used to require massive human resources. Adding to the complexity, some of the most sought-after platforms, such as Twitter and Reddit, publish text data and multimedia content such as images or videos, making data extraction even more challenging. Fortunately, researchers have made similar advances in other areas of AI, such as audio, image, and video processing, unlocking the potential of information sources that were previously too labor-intensive to process.

7 Evaluating Alternative Data

Determining the value of an alternative dataset relies heavily on the individual firm using it, their purpose, and the kind of information they are hoping to gain. A current roadblock to the use of alternative data is management. In a 2017 survey by Greenwich Associates, nearly 40% of hedge fund managers were not convinced of the value of alternative data (Putting Alternative Data to Use in Financial Markets, 2017). One of the main arguments in favor of the value of alternative data is demonstrated by outlining the risks of late adoption of alternative data, which is carried out below.

The three main risks of late adoption of alternative data outlined by a Deloitte Industry Report are positioning risk, execution risk, and consequence risk. Positioning risk highlights how markets become more efficient through the wide-scale adoption of alternative data. Price signals acted upon by informed alternative data users leave non-informed participants at a disadvantage. Execution risk outlines how firms adopting and developing alternative data into their current decision-making process have the processes and internal expertise they need to stay ahead of the competition. A key example of this is the hiring of alternate data analysis teams. Firms that decide to "wait and see" may end up without the talent they need to remain competitive. Finally, consequence risk explains the "strategic risk" that may occur to the company if alternative data is not employed, and the company loses out to peers in terms of innovation. As the report outlines, "Today's innovation could be tomorrow's requirement" (*Alternative Data for Investment Decisions: Today's Innovation Could Be Tomorrow's Requirement*, 2017, p. 1).

As a result of all of these factors, evaluating a dataset is no trivial task. The value of a new dataset should be measured based on the value of the information gained over and above the data already in use. The incremental information can be used by financial actors to improve financial decision-making and the resulting profits. A common method to evaluate value in this context is back-testing. This process consists of playing out alternative historical scenarios of the investment outcome if the decision-makers had access to the new dataset. One major limitation to this approach is that alternative datasets are recent and lack the historical records needed for proper back-testing. While back-testing might provide some insights into value, the competitive environment and the continuous adoption of new datasets by competitors also limit the power of back-testing. Paper profits seen in back-testing may not reflect future profits if competitors are simultaneously also adopting new datasets.

8 Conclusion

This chapter presented an overview of the alternative data landscape as it stands today. Driven by digitalization and an ever-mounting number of data points, alternative data now deserves a place in the decision-making process more than ever. It has never been more affordable to store vast amounts of data and acquire the computing power needed to process these insights. The use cases of alternative data are extensive, and the main applications covered in this chapter reflect those most used at this present time. Currently, alternative data in all forms, including text, social media-focused, transaction, and satellite imagery, give real insights into investment opportunities. In 2016, the State Street Chairman and CEO mentioned their ability to conduct more in-depth investigations of inflation by tracking price fluctuations on several million items; these insights allowed State Street to identify price shocks faster and pinpoint those segments most impacted (Ryan, 2016). This example shows how financial sectors that are heavily reliant on information intake will benefit the most from adopting alternative data into their workflows.

Regarding adoption, hedge funds have been the innovators in this sector for a long while and are now being joined by "early adopters" such as aggressive long-only managers and private equity firms. Investment managers vary in their approach, but the data shows that those who are more tech-savvy are becoming the majority (Dannemiller & Henry, 2018). Coincidentally, the sectors that benefit the most from alternative data adoption also face the most significant risks of late adoption, as the advantages gained by investment managers who utilize alternative data are beginning to increase. A real-life example of this advantage can be seen even in classic forecasting of earnings reports. In 2018, Tesla Motors had yet to turn a profit in its history. Months before their Q2 disclosure, CEO Elon Musk vowed to increase production by working around the clock in hopes of meeting production targets. Thasos Group collected geolocation data from over 1,000 apps, marking smartphone pings like beacons on a map. Thasos recorded a 30% increase in the overnight shift from these smartphone beacons, Tesla recorded a "rare quarterly profit," and the number of completed Model 3s doubled. Thasos Group is rolling out these data insights via Bloomberg's terminal, showing that if a firm has not employed alternative data on its own accord, it will soon become part of the default data suite in finance (Dezember, 2018).

For reasons evident in this chapter, describing all of the alternative data available today or predicting which new type of data will be dominant in the future is not possible. However, the lessons and guiding principles outlined in this chapter should help financial analysts and decision-makers to plan and evaluate the costs and benefits of embracing alternative data. As the cost of acquiring and processing data is still declining, we can expect the adoption of alternative data to accelerate. Because advances in machine learning allow the automatic processing of complex unstructured data, we can expect future growth to come from more complex sources such as text, audio, images, and videos. The avenues for gathering alternative datasets are vast. This chapter covers several text data applications, including social media, credit card or e-commerce transaction data, satellite data, and weather data. Although they all have different use cases, the purpose of these datasets is to generate new ideas, present additional perspectives, gain information, and reduce uncertainty. The cost of implementing an in-house team may discourage many from the use of alternative data. However, a burgeoning industry of intermediary data providers and aggregators allows firms to access alternative data at a fraction of the cost. This growth in providers and aggregators may shift the narrative from "is alternative data worth the cost?" to "can we afford not to use it?".

References

Alternative data for investment decisions: Today's innovation could be tomorrow's requirement. (2017). Deloitte Center for Financial Services. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/financial-services/us-fsi-dcfs-alternative-data-for-investment-decisions.pdf

Antweiler, W., & Frank, M. Z. (2004). Is all That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance, 59*(3), 1259–1294. https://doi.org/10.1111/j.1540-6261.2004.00662.x

Baker, S., & Kueng, L. (2021). *Household financial transaction data* (No. w29027, p. w29027). National Bureau of Economic Research. https://doi.org/10.3386/w29027

Bennett, H. M., Granik, G., & Sacerdoti, A. J. (2010). *Repeal of dodd-frank confidentiality protection for SEC: What investment advisers lost and what remains*. https://www.hflawreport.com/2541071/repeal-of-doddfrank-confidentiality-protection-for-sec-what-investment-advisers-lost-and-what-remains.thtml?1

Blankespoor, E., Miller, G. S., & White, H. D. (2013). The Role of Dissemination in Market Liquidity: Evidence from Firms' Use of Twitter™. *The Accounting Review, 89*(1), 79–112. https://doi.org/10.2308/accr-50576

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(January), 993–1022.

Bodas-Sagi, D. J., Garcia, J. R., Rodrigo, T., Ruiz de Aguirre, P., Ulloa, C. A., Murillo Arias, J., de Dios Romero Palop, J., Valero Lapaz, H., & Pacce, M. (2019). *Measuring retail trade using card transactional data* (SSRN Scholarly Paper ID 3414182). Social Science Research Network. https://papers.ssrn.com/abstract=3414182

Butcher, S. (2022, February 17). *Alternative data jobs in finance are going crazy*. EFinancial Careers. https://www.efinancialcareers.co.uk/news/2022/02/alternative-data-jobs

Cohan, P. (2018, July 22). *Mastercard, AmEx and envestnet profit from $400M business Of Selling transaction data*. Forbes. https://www.forbes.com/sites/petercohan/2018/07/22/mastercard-amex-and-envestnet-profit-from-400m-business-of-selling-transaction-data/

Cookson, J. A., & Niessner, M. (2020). Why Don't We Agree? Evidence from a Social Network of Investors. *The Journal of Finance, 75*(1), 173–228. https://doi.org/10.1111/jofi.12852

Dannemiller, D., & Henry, P. (2018). *Alternative data adoption in investment management: Infocus*. Deloitte. https://www2.deloitte.com/us/en/pages/financial-services/articles/infocus-adopting-alternative-data-investing.html

Denev, A., & Amen, S. (2020, June). *The book of alternative data: A guide for investors, traders and risk managers*. Wiley. https://www.wiley.com/en-us/The+Book+of+Alternative+Data%3A+A+Guide+for+Investors%2C+Traders+and+Risk+Managers-p-9781119601807

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:* https://arxiv.org/abs/1810.04805

Dezember, R. (2018, November 2). Your smartphone's location data is worth big money to wall street. *Wall Street Journal*. https://www.wsj.com/articles/your-smartphones-location-data-is-worth-big-money-to-wall-street-1541131260

Drake, M. S., Roulstone, D. T., & Thornock, J. R. (2012). Investor Information Demand: Evidence from Google Searches Around Earnings Announcements. *Journal of Accounting Research, 50*(4), 1001–1040. https://doi.org/10.1111/j.1475-679X.2012.00443.x

Duz Tan, S., & Tas, O. (2021). Social Media Sentiment in International Stock Returns and Trading Activity. *Journal of Behavioral Finance, 22*(2), 221–234. https://doi.org/10.1080/15427560.2020.1772261

EY Global Alternative Fund Survey. (2019). Ernst & Young Global Limited. https://assets.ey.com/content/dam/ey-sites/ey-com/en_gl/topics/wealth-and-asset-management/ey-global-alts-fund-survey-final.pdf

Fleming-Williams, M. (2021). *The Jason hills episode* (No. 64) [Mp4]. Retrieved January 2, 2022, from https://open.spotify.com/episode/66pRJqfxMCa764KHBau2B4?si=d2e3a4ddd0eb4aae&nd=1

Fleming-Williams, M. (2020). *What is alternative data?* (No. 01) [Mp4]. Retrieved January 2, 2022, from https://open.spotify.com/episode/4Y7KY5f5TDEIKndVyJPKac

Freedom of Information Act: Federal Agencies' Recent Implementation Efforts. (2020). Freedom of information act: Federal agencies' recent implementation efforts. https://www.gao.gov/products/gao-20-406r

Gargano, A., Rossi, A. G., & Wermers, R. (2017). The Freedom of Information Act and the Race Toward Information Acquisition. *The Review of Financial Studies, 30*(6), 2179–2228.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature, 57*(3), 535–574. https://doi.org/10.1257/jel.20181020

Global Social Media Stats. (2022, January). Data reportal—Global digital insights. https://datareportal.com/social-media-users

Goldman Sachs and Citi Make $20 Million Venture Bet That Private Equity Wants to Your Credit Card Info. (2019, February 12). Fortune. https://fortune.com/2019/02/12/goldman-sachs-citi-second-measure-alternative-data/

Green, T. C., Huang, R., Wen, Q., & Zhou, D. (2019). Crowdsourced employer reviews and stock returns. *Journal of Financial Economics, 134*(1), 236–251. https://doi.org/10.1016/j.jfineco.2019.03.012

Gupta, A., Dengre, V., Kheruwala, H. A., & Shah, M. (2020). Comprehensive review of text-mining applications in finance. *Financial Innovation, 6*(1), 39. https://doi.org/10.1186/s40854-020-00205-1

Gutiérrez, E., Lourie, B., Nekrasov, A., & Shevlin, T. (2020). Are Online Job Postings Informative to Investors? *Management Science, 66*(7), 3133–3141. https://doi.org/10.1287/mnsc.2019.3450

İç, S., Kahyaoğlu, M. B., & Odabaş, D. (2014). The Effects Of Weather On Investor Behavior: A Study On Individual Turkish Stock Market Investors. *Journal of Business Economics and Finance, 3*(2), 191–206.

Industry stats. (2019). Alternative data. https://alternativedata.org/stats/

Liberti, J. M., & Petersen, M. A. (2019). Information: Hard and Soft. *The Review of Corporate Finance Studies, 8*(1), 1–41. https://doi.org/10.1093/rcfs/cfy009

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65.

Marr, B. (2018). *How much data do we create every day? The mind-blowing stats everyone should read*. Forbes. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/

Monk, A., Prins, M., & Rook, D. (2018). Rethinking Alternative Data in Institutional Investment (SSRN Scholarly Paper ID 3193805). *Social Science Research Network*. https://doi.org/10.2139/ssrn.3193805

Mullins, B. & Weaver, C. (2013, September 23). Open-government laws fuel hedge-fund profits. *Wall Street Journal*. https://www.wsj.com/articles/opengovernment-laws-fuel-hedgefund-profits-1379905298

Personal Finance Is a Hit on TikTok. (2022, January 29). *The Economist*. https://www.economist.com/finance-and-economics/2022/01/29/personal-finance-is-a-hit-on-tiktok

Predictive Modeling Using Transactional Data. (2010). Capgemini. https://www.capgemini.com/resources/predictive-modeling-using-transactional-data/

Putting Alternative Data to Use in Financial Markets. (2017). Greenwich. https://www.greenwich.com/institutional-investing/putting-alternative-data-use-financial-markets

Ryan, D. (2016, September 1). *State street's Joseph Hooley advises investors to explore diverse datasets*. Data Strategy. https://dsg.eaglealpha.com/state-streets-joseph-hooley-advises-investors-to-explore-diverse-datasets/

Sager, M. (2020). *Freedom of information act: Federal agencies' recent implementation efforts*. US Government Accountability Office. https://www.gao.gov/products/gao-20-406r

SEC.gov | SEC Staff Releases Report on Equity and Options Market Structure Conditions in Early 2021. (2021, October 14). https://www.sec.gov/news/press-release/2021-212

Sharma, V. (2018, December 10). *How did digital storage (effectively) become free*. USC Marshall. https://www.marshall.usc.edu/blog/how-did-digital-storage-effectively-become-free

Talib, R., Kashif, M., Ayesha, S., & Fatima, F. (2016). Text mining: techniques, applications and issues. *International Journal of Advanced Computer Science and Applications*, *7*(11). https://doi.org/10.14569/IJACSA.2016.071153

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.

Tomaselli, O., & Kilduff, E. (2018). *Staffing for the data deluge*. http://www.optionsgroup.com/Strategies-Solutions/Thought-Leadership/Staffing-For-Data-Deluge

Tumarkin, R., & Whitelaw, R. F. (2001). News or Noise? Internet Postings and Stock Prices. *Financial Analysts Journal, 57*(3), 41–51. https://doi.org/10.2469/faj.v57.n3.2449

Wagner, K. (2021, July 14). Facebook users said no to tracking. Now Advertisers are Panicking. https://www.bloomberg.com/news/articles/2021-07-14/facebook-fb-advertisers-impacted-by-apple-aapl-privacy-ios-14-changes

Wigglesworth, R. (2018, January 29). Asset management's fight for 'alternative data' analysts heats up. *Financial Times*. https://www.ft.com/content/2f454550-02c8-11e8-9650-9c0ad2d7c5b5

An Algorithmic Trading Strategy to Balance Profitability and Risk

Guillermo Peña

1 Introduction

Big data's development inspired a complete transformation of many disciplines, and finance is no exception. Due to the vast quantity of information that computers and the internet both keep and process, as well as consequent statistical advances, there are many long panel data available with rich and diverse information. This shift offers a great opportunity for developing systematic trading algorithms that allow for more efficient investment with higher flexibility and profits, as well as increased speed and enhanced knowledge. Nonetheless, it is also essential to balance the profit-risk trade-off. With this in mind, this chapter revises current algorithmic trading (AT) methodologies and builds on existing algorithms in the literature. In doing so, it uses the methodology of designing algorithms based on long, historical datasets. This chapter employs the full

All the Tables and Figures are author originated.

G. Peña (✉)
University of Zaragoza, Zaragoza, Spain
e-mail: gpena@unizar.es

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_3

sample of exports and imports from the World Bank database, with data from more than 200 countries from 1960 to 2018.

Algorithmic trading (AT) is one of the most widely used techniques for choosing a preferred portfolio. AT generally refers to the application of sophisticated or complex quantitative algorithms to automatically determine whether to buy or sell in the financial trading cycle of markets (Hu et al., 2015). This activity includes pre-trade analysis (using data analysis), the generation of trading signals (through recommendations related to buying and selling), and finally, trade execution (by ordering management). Since the 1990s, this analysis has been based on discovering rules after data-mining and analyzing big data. Despite the complexity underpinning both markets and most AT methods (Hilbert & Darmon, 2020), such as the development of artificial intelligence or machine learning, there is a rising demand for practical solutions and simple methods that manage this array of techniques (Hansen, 2020). Indicators developed via observing historical data and with the tools that big data provides can meet this demand.

An indicator is a quantitative signal that provides specific information to the user. Well-known economic and financial indicators are the Gini index for measuring income inequality, the Herfindahl–Hirschman index for measuring market power, the Altman Z-score for predicting bankruptcy, and the net interest margin as a measure of banking profitability. This chapter provides a combination of banking default predictability and profitability inspired by the last two indicators for ensuring an investment with relatively high profitability and reasonably low default risk. Traditionally, there has been a trade-off between profitability and risk. For instance, investing in risker stocks usually leads to higher short-term profits. The main reason generally considered is that the higher profitability is already paying the investor for possible future defaults. Though AT strategies allow for higher short-term profits, it is also worth considering the default risk these strategies present in the long run. The fact remains that despite the increase in profitability and better control over volatility that businesses that follow AT strategies have (Virgilio, 2019), there are some potential disadvantages for other traders in the financial markets (Yan, 2019). One such disadvantage is the presence of a commonly used technical employment of AT (based on trends and above-average data) rather than following each enterprise's economic fundamentals (as the real economic data or financial accounts). Together with the high profit-low risk dichotomy, such trade-offs are some of the reasons behind the search for an AT strategy in which there is an accurate equilibrium between returns and risk. This chapter tackles this issue by using a proxy of the economic fundamentals, which includes the proposed indicators that are checked as reasonable measures of profitability and bankruptcy suggested by Ho and Saunders (1981) and Altman (1968). Several indicators are proposed by the author and developed to address the trade-off issue. This chapter also checks and cross-references these indicators with trading commerce data to ensure their approximate economic fundamentals and applicability. This study uses fictional examples of financial operations to analyze which portfolios would be chosen when employing the "balanced investment indicator" (BII).

Moreover, this chapter's objective is to propose a sophisticated yet straightforward AT strategy to balance market returns and low economic risks. The study presented meets this target by formulating an indicator that weights the portfolios by a midpoint between a modified bid-ask spread to measure default risk and a modified mid-price as a measure of profitability. With this purpose in mind, the following section explains the concept of AT and provides the main methods, strategies, and means of analysis that AT uses. This section also presents a general overview of financial and capital markets. The rest of the chapter is outlined as follows: Sect. 3 proposes indicators and rules useful to AT strategies; Sects. 4, 5, and 6 empirically check these indicators and rules with real data from trading countries, stocks, and the prices of stock exchanges, in addition to data from a fictional comparative example. Moreover, Sect. 6 also discusses the importance of big data in achieving an accurate choice of portfolios through the proposed AT strategy. Finally, Sect. 7 concludes the chapter.

2 Algorithmic Trading: Concept, Methods, and Influence

Many frequently used AT strategies use historical prices from stock markets as a tool to make buying and selling decisions in financial markets. Pricope (2021) states the importance of buying at relatively low prices ("bullish signals," where bullish refers to market environments with a lower number of purchases than sells) and of selling when there are high prices. The primary goal of AT is to maximize returns by exploiting the volatility of prices, but also considering economic factors in order to avoid the inherent risk for investment.

Other sources of information for AT include financial and economic data, such as financial accounts and business statements, or macroeconomic data that provides contextual and long-run economic information such as the gross domestic product (GDP), public debt, and the inflation rate of the country where the business is settled. Financial news can also be used as a source of information in this context (Feuerriegel & Prendinger, 2016; Hafezi et al., 2015).

Moreover, other ways of establishing trading algorithms are based on theories or are empirically based (for a review of different methods, see Huang et al., 2019; Treleaven et al., 2013). Theories derived from academic fields such as economics, social sciences, physics, and econophysics form the basis for the design of AT strategies and algorithms. In some cases, empirical evidence is what guides the design of an indicator based on trends, as with the technical analysis or econometric and statistical models that confirm the theoretical expectancies. Some strategies consist of buying the financial product with the highest profitability, while others focus on the overpriced or underpriced portfolios relative to the historical trend or obtaining capital gains after relatively long periods of increase in the price (Dunis et al., 2004; Kaufman, 2005). Additionally, the natural world inspires other techniques such as the evolutionary computation of algorithms or swarm intelligence to optimize profits (Hu et al., 2015). Biological evolution and the process of natural selection and mutations inform these algorithms, which are then adapted to the selection of portfolios.

As previously mentioned, the latest research in this field by Hu et al. (2015) states that there are three analytical methods for assessing portfolios:

1. **Fundamental**: based on the characteristics and the profitability of the business, market, and socio-economic context,
2. **Technical**: based on the momentum of the portfolio in the market, as the trend of the movements of the stock prices and volumes, which reflect the information regarding the stock, or
3. **Blending analysis**: a mixture of fundamental and technical analyses.

Practitioners of banks and financial institutions developed a traditional case of mixing strategies. They usually consider the immediate price of the stocks with respect to the *mid-price* (the arithmetic average between purchases and sells, as a measure of technical analysis) in addition to the financial accounts of the business behind the stock (the projections of future growth, organic growth, as fundamental analysis). Through AT, big data can complement these considerations. For instance, using techniques specific to the science of big data allows for the use of a large amount of information regarding stock prices and trends to conduct technical analysis. Additionally, big data techniques could also be used with large amounts of information from economic data and financial accounts to conduct fundamental analysis. Both of these analyses to be used in AT would be very difficult to conduct without big data tools.

The widespread use of AT can impact trading itself and the economy more generally. Virgilio (2019) suggests the profound influence of AT on the microstructure of financial markets, as illustrated by its use in providing liquidity or for controlling volatility, which can damage the strategies of other traders. The author even considers whether using these techniques can reduce the agents' profits. Furthermore, Yan (2019) finds real effects on the financial statements of businesses, such as the deterioration of the businesses' accounts or the mismatching between corporate investment and stock prices, when traders implement these methods. This chapter attempts to mitigate the trade-off above by balancing returns and moderating risk using empirical evidence. The proposed strategy, to be outlined in the next section, is based on technical analysis using the historical trends of the prices and fundamental analysis using the real economic data of historical trading patterns among different countries. The major gap in the literature this chapter addresses is finding an AT strategy that yields an appropriate equilibrium between earning income and managing risk.

3 Proposed AT Strategy

To start, it is essential to distinguish between market default risk and the price associated with a financial product's return or profitability. If the product's profitability is high, the price (measured by the mid-price) is usually high as well. The arithmetic average between the ask (the price that sellers initially intend to offer, as related to supply) and the bid (the price at which the buyer intends to purchase, as related to demand) expresses the mid-price:

$$\mathrm{MP} = \frac{\mathrm{ask} + \mathrm{bid}}{2}. \tag{1}$$

On the other hand, the market risk of trading can be higher if the quoted spread widens. The quoted spread measures the gap between supply and demand through the following formula:

$$\mathrm{QS} = \frac{\mathrm{ask} - \mathrm{bid}}{\mathrm{MP}}. \tag{2}$$

Per López-Laborda and Peña's (2018) and Peña's (2020, 2021) findings, this chapter proposes the use of a modified quoted spread, called the "mobile-ratio" (MR), which has optimal operative behavior and additional applications in public finance:

$$\mathrm{MR} = \frac{\mathrm{QS}}{2}. \tag{3}$$

Additionally, building on López-Laborda and Peña (2018) and Peña's (2020) research, this chapter proposes an additional indicator with a high correlation and specific risk-free financial products, called the "pure price" (PP). The same units are used as the previous MR indicator by dividing it by twice the mid-price (MP):

$$\mathrm{PP} = \frac{\mathrm{ask} \cdot \mathrm{bid}}{\mathrm{MP}}. \tag{4}$$

The last indicator used in this chapter is called the "indicator of unitary profit" (IUP), which is formulated below:

$$\mathrm{IUP} = \frac{\mathrm{ask} \cdot \mathrm{bid}}{2 \cdot \mathrm{MP}^2}. \tag{5}$$

This chapter may consider the IUP as negatively correlated with the return, which is further explored by the author in Sect. 4. Regarding the MP and IUP indicators, they can be expressed as functions of a and b, where $a = \mathrm{ask}/(\mathrm{ask} + \mathrm{bid})$ and $b = \mathrm{bid}/(\mathrm{ask} + \mathrm{bid})$:

$$\mathrm{MR} = \max\{a, b\} - \min\{a, b\},\ \ \mathrm{IUP} = 2 \cdot a \cdot b. \tag{6}$$

These two indicators are bounded as the maximum value of a and b is one, and the minimum is zero. Therefore, it follows that the maximum value for the MR indicator is one, while with zero transaction frictions, where $a = b = 0.5$, the minimum is zero. Additionally, the IUP indicator is bounded between 0 and 0.5, depending on the absolute value of a and b, and so the lowest MP value gives the highest IUP value: $2*0.5*0.5 = 0.5$. Therefore, the lowest value of the IUP indicator is given when the MP is at its maximum: $2*1*0 = 0$, and so the best comparability of the indicators is achieved by normalizing the MR between 0 and 0.5, which allows for division by 2.

Additionally, the trade-off between risk and return (Adrian et al., 2019; Aggarwal & Samwick, 1999) is addressed in this chapter by establishing a unique indicator in which the sum of the two normalized indicators balances and measures the market risks and the returns. This unique indicator, called the balanced investment indicator (BII), is outlined below:

$$\mathrm{BII} = \mathrm{IUP} + \frac{\text{Quoted Spread}}{4}. \tag{7}$$

As a result, the potential values of this indicator fall between zero and one.

Furthermore, the proposed AT strategies are best used in combination with panel data, which is defined as a large amount of time-series data. Consequently, the MR and IUP value of an individual portfolio or value i and time t are as follows:

$$\begin{aligned} \mathrm{MR}_i &= \frac{\text{Quoted Spread}_i}{2} = \sum_{t=1}^{T} \frac{\mathrm{Ask}_{it} - \mathrm{Bid}_{it}}{\mathrm{Ask}_{it} + \mathrm{Bid}_{it}}, \\ \mathrm{IUP} &= \sum_{t=1}^{T} \frac{2 \cdot \mathrm{Ask}_{it} \cdot \mathrm{Bid}_{it}}{(\mathrm{Ask}_{it} + \mathrm{Bid}_{it})^2}. \end{aligned} \tag{8}$$

Additionally, the BII of a specific value or portfolio is defined as:

$$\mathrm{BII}_i = \mathrm{IUP}_i + \frac{\mathrm{MR}_i}{2} = \mathrm{IUP}_i + \frac{\text{Quoted Spread}_i}{4}. \tag{9}$$

Finally, and considering BII without the subscript of the average of the BII of the full sample, the proposed algorithm of balanced investment

(ABI) is as follows:

$$\text{ABI}:\begin{cases}\text{BII}_i - \text{BII} > \text{Comfort top} \rightarrow \boldsymbol{Buy}\\ \text{BII}_i - \text{BII} \in [\text{Comfort bottom},\ \text{Comfort top}] \rightarrow \boldsymbol{Keep}\\ \text{BII}_i - \text{BII} < \text{Comfort bottom} \rightarrow \boldsymbol{Sell}\end{cases} \quad (10)$$

The observation of historical, real-economy-based data forms the basis of this algorithm and is based on technical analysis but with some fundamental criteria. With a "comfort" top (since that is the value that the investor purchases) and with a "comfort" bottom (up to which the investor sells), the investor using this AT algorithm may choose both the bottom and top when making investment decisions. The examples presented throughout this chapter assume that the comfort top equals the comfort bottom, but other alternatives exist. For instance, another option is considering the comfort top and bottom as a r and s, multiples of BII respectively, where $r = 1.5$ and $s = 0.5$, instead of $r = s = 0$ as used here.

4 Empirical Evidence of Proposed AT Strategies

To conduct the empirical analysis in this section, this chapter initially uses an available historical dataset comprising economic trading variables. This dataset is used to calibrate the suitability of the proposed indicators that form the basis of the algorithm. Moreover, the chapter uses the complete sample of more than 200 countries from the World Bank for the years 1960 to 2018 for the variables "Commercial service exports" (X) and "Commercial service imports" (M). In this example, the maximum and minimum replace the bid and ask, respectively, with X and M for each country and period. This example employs a sufficiently comprehensive dataset with several periods, but in the rest of the examples, this study uses only one period for two reasons. The first reason is the difficulty of obtaining temporal series of bids and asks, while the second reason, connected to the first, is that using larger samples was necessary for the statistical analysis to draw more thorough conclusions from the data. As discussed later in this chapter, big data methods have the potential to be tremendously valuable to the algorithm presented by being able to find the most accurate threshold of investment, and because of the allowed comparability between periods and portfolios.

4.1 Empirical Analysis

Figure 1 shows the correlations between the IUP, MR, and BII indicators for the sample of countries.

As can be seen by the reader in Fig. 1, an inverted U in the relationship of the BII indicator with respect to IUP and MR indicators is observed. The inverted U shape of both relationships, with a maximum BII in the middle value (roughly) of the indicators, reflects a balance (or equilibrium) in the trade-off of return-risk. Nonetheless, it is ensured by the self-working of the proposed indicators that the value with the highest BII value does not usually coincide with the country associated with the average nor median of the IUP or MR indicators.

Figure 2 shows the relationship between the gross domestic product (GDP) growth rate in 2018 and the BII indicator for the studied countries. This relationship is negative for the full sample (left) but is positive for the 15 countries with the highest indicator value (right). This result again reflects the equilibrium between profitability and risk in those countries, considering the absence of risk for the full sample, but higher profitability for the chosen countries.

Furthermore, Table 1 shows the average obtained for the MR, IUP, and BII indicators for countries, stocks, and stock exchange samples. These findings reveal that, while the BII indicator value is close for

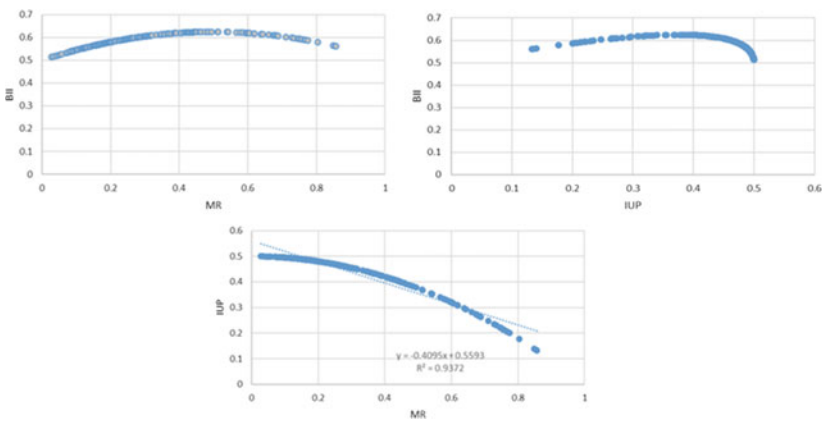


Fig. 1 Correlations between the three main indicators: indicator of unitary profit (IUP), mobile-ratio (MR), and balanced investment indicator (BII)

the three samples, there is a difference in the MR and IUP indicators when the data is expressed in an aggregated way (countries and stock exchanges) compared to when it is individual (stocks).

Moreover, Tables 2, 3, and 4 show the countries with the highest BII, lowest MR, and highest MR values. These obtained values represent, respectively, the countries with the maximum value on the balance between profitability and risk (countries such as Saudi Arabia or Bangladesh, with a high or a rising level of development, but not an excessive risk), countries with the lowest risk of default (countries such as the Netherlands or Norway), and finally, countries with the highest risk (including Libya and the Democratic Republic of Congo).

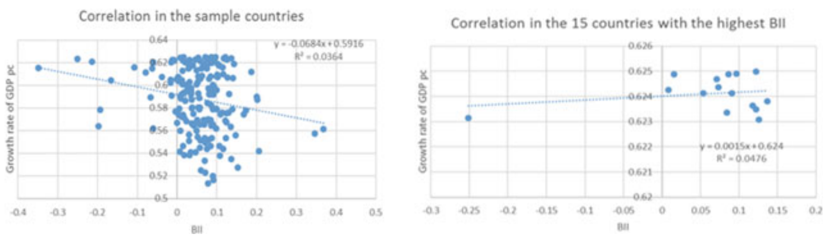


Fig. 2 Correlations between economic activity and the balanced investment indicator (BII)

Table 1 Average value of the indicators for the three samples

Average	*MR*	*IUP*	*BII*
Countries	0.32096299	0.427828	0.58841492
Stocks	0.50034822	0.33919124	0.58936534
Stock Exchanges	0.37655987	0.40111053	0.58939047

Table 2 Countries with the highest balanced investment indicator (BII) in the sample

Countries with highest BII	*MR*	*IUP*	*BII*
Saudi Arabia	0.4948	0.3776	0.625
Eritrea	0.5114	0.3692	0.6249
Cote d'Ivoire	0.5142	0.3678	0.6249
Zimbabwe	0.4851	0.3824	0.6249
Bangladesh	0.4846	0.3826	0.6249

Table 3 Countries with the lowest mobile-ratio (MR) in the sample

Countries with lowest MR	*MR*	*IUP*	*BII*
Belgium	0.0278	0.4996	0.5135
Netherlands	0.0335	0.4994	0.5162
Norway	0.0412	0.4992	0.5197
Italy	0.0481	0.4988	0.5229
Australia	0.0529	0.4986	0.5251

Table 4 Countries with the highest mobile-ratio (MR) in the sample

Countries with highest MR	*RM*	*IUP*	*BII*
Libya	0.8572	0.1326	0.5612
Burundi	0.8555	0.134	0.5618
Angola	0.849	0.1396	0.5641
Chad	0.8039	0.1769	0.5788
Congo, Rep	0.7742	0.2003	0.5874

5 Comparison of Proposed AT Strategy with Other Benchmarks

In this section, the proposed indicators have been adapted to compare with current benchmarks due to the available information on the stocks from the IBEX-35 of the Madrid stock exchange. The available information is the maximum, minimum, and closing price of the stocks on January 14, 2022, for this exchange. The BII indicator uses the same MR and IUP indicators as in Eq. (6), but in this case, the a and b of Eq. (5) are obtained as the difference between the maximum price and the closing price, and the closing price minus the minimum price.

5.1 Empirical Evidence of AT Strategy Using IBEX-35 Exchange

The results presented in Table 5 show that the BII indicator allows us to find the stock Cellnex as one of the four best stocks in the IBEX-35 exchange. These results echo those of a recent report from Bank of America in which the firm situated this stock, jointly with Telefonica, as one of five European telecoms with the highest potential of revaluation in 2022 (*Expansión*, 2022). Additionally, the Spanish financial newspaper *Expansion* considers Cellnex as one of the stocks that the IBEX-35 favors this year, with expected returns of 60% (*Expansión*, 2022).

Table 5 Stocks with the highest balanced investment indicator (BII) on the IBEX-35 exchange

Stocks in IBEX 35 with highest BII	*MR*	*IUP*	*BII*
IAG	0.493	0.3785	0.625
INM.COLONIAL	0.4857	0.382	0.6249
BANKINTER	0.5217	0.3639	0.6248
CELLNEX	0.5227	0.3634	0.6247

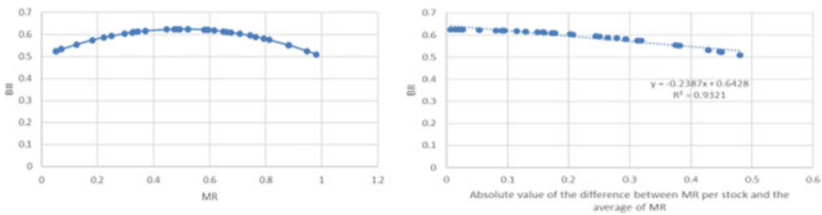


Fig. 3 Relationship between the mobile-ratio (MR) and balanced investment indicator (BII) (left) and between the deviation from the MR average and BII (right)

Moreover, Fig. 3 shows the balance in the trade-off between high risk (high MR) and low risk (low MR), as that which achieves the highest BII indicator score. This phenomenon is observed in the highly explanatory negative relationship between the BII indicator and the deviation of the MR indicator from the average. A higher deviation from the MR's average means a lower BII indicator value, but this relationship is not strictly linear. This relationship is slightly positive or null for data with the lowest deviation.

5.2 Evidence of AT Strategy Using Fictional Data and Other AT Strategies

In addition to showing the usage of the BII indicator with stock benchmarks, a fictional exercise proposed herein compares the indicator with other benchmarks traditionally used in AT. Concretely, and starting from an example of the bids and asks of hypothetical prices of stocks, three methodologies are compared: first, Peña's (2020) proposal; second, an example of usual practice in many financial institutions (with slight differences sometimes among them); and finally, the investment decision that this chapter presents using the proposed BII indicator. The previous methodology has been used because it serves as a close comparison to the current proposal. It is also relevant to compare the contribution of this chapter with the usual practices of the financial sector to specify the benefits of the proposed strategy.

Based on the assumption of the bottom and top equality of Eq. (10), Tables 6, 7, and 8 present the results of the following fictional exercise. The example considers five portfolios with their respective asks (A) and bids (B) in monetary units. Specifically, Table 6 provides the same example used for the three methodologies by considering only the price specifications of the five financial products with different asks and bids, *ceteris paribus*, with the products called "a-e." According to Peña's (2020) proposal, the decision to invest or not depends on a K-indicator as the arithmetic average between the supply and demand defined as the spread (difference between the bid and ask) and the pure price (twice the product of the bid and ask divided by the sum of both). Selected portfolios are those with a K-indicator lower than the average of the portfolios (e.g., portfolios b, d, and e).

Table 7 considers an example of investing by the usual practices used in finance, which, in this case, is the investable portfolios with the mid-price (arithmetic average of the prices) lower than the average. Therefore, the investable portfolios are d and e.

In contrast, Table 8 provides the results in which the MR indicator and "delta" are obtained, respectively, as the spread and pure price, but divided by the sum of bid and ask. The BII and the MR indicators are

Table 6 Investing-selling decisions according to Peña's (2020) proposal

Product	*A*	*B*	*VA (Spread)*	*Pure price*	*K*	*Invest?*	*Sell?*
a	0.6	0.4	0.2	0.48	0.34	No	Yes
b	0.5	0.3	0.2	0.375	0.2875	Yes	No
c	0.8	0.2	0.6	0.32	0.46	No	Yes
d	0.4	0.1	0.3	0.16	0.23	Yes	No
e	0.3	0.1	0.2	0.15	0.175	Yes	No
				Average	0.2985		

Note VA stands for Value Added

obtained by dividing by two plus the delta. In this case, the investment decision hinges on whether the BII indicator is higher than the portfolio's average. This strategy results in investing in portfolios c, d, and e.

Next, Table 9 summarizes the investment decisions for the three methodologies according to the results that Tables 6, 7, and 8 present. Portfolios d and e are investable per the three approaches, with those two portfolios being the unique investable portfolios by the usual methodology. There are also investable portfolios b per Peña's (2020) proposal and c for the current proposal, adding an alpha of value to the trading in both cases. In the first case, the alpha is the midpoint between supply and demand, and in the second case, it is the balance between profitability and low risk.

This simple example provides insight into approaches different from those proposed in this chapter and highlights the potential alpha (differentiation in trading from other traders that provides a plus of profitability) of the contribution of this chapter.

Table 7 Investing-selling decisions according to current practices

Product	*A*	*B*	*Mid-price*	*Invest?*	*Sell?*
a	0.6	0.4	0.5	No	Yes
b	0.5	0.3	0.4	No	Yes
c	0.8	0.2	0.5	No	Yes
d	0.4	0.1	0.25	Yes	No
e	0.3	0.1	0.2	Yes	No
		Average	0.37		

Table 8 Investing-selling decisions according to the current proposal

Product	*A*	*B*	*VA (Spr.)*	*Pure price*	*MR*	*Delta*	*BII*	*Invest?*	*Sell?*
a	0.6	0.4	0.2	0.48	0.2	0.48	0.58	No	Yes
b	0.5	0.3	0.2	0.375	0.25	0.46875	0.594	No	Yes
c	0.8	0.2	0.6	0.32	0.6	0.32	0.62	Yes	No
d	0.4	0.1	0.3	0.16	0.6	0.32	0.62	Yes	No
e	0.3	0.1	0.2	0.15	0.5	0.375	0.625	Yes	No
						Average	0.608		

Note Prod. and Spr. Mean Product and Spread, respectively. A is the ask price and B is the bid price

Table 9 Summary of investing-selling decisions using the three different methods

Product	*Current proposal*		*Peña's (2020) proposal*		*Usual way*		*Comments*
	Invest?	*Sell?*	*Invest?*	*Sell?*	*Invest?*	*Sell?*	
a	No	Yes	No	Yes	No	Yes	
b	*No*	*Yes*	*Yes*	*No*	*No*	*Yes*	*Opportunity for investing with alpha for the* Peña (2020) *proposal*
c	**Yes**	**No**	**No**	**Yes**	**No**	**Yes**	**Opportunity for investing with alpha for the current proposal**
d	Yes	No	Yes	No	Yes	No	
e	Yes	No	Yes	No	Yes	No	

6 Discussion and Applicability of Big Data to Proposed AT Strategy

It is worth highlighting the relevance of big data in the application of this approach, and AT in general for three main reasons. First, we can use big data to compare portfolios, selecting the most optimal option and providing the average in the market. Second, big data can assist in extrapolating from other variables with a high abundance of associated data and by using machine learning or artificial intelligence techniques (neuronal approaches, for instance) to choose the thresholds for buying/keeping/selling, the bottom and top from Eq. (9). Lastly, big data can identify the best choice for investing, not only for making comparisons between options or with respect to the average or threshold, as in points one and two. In this case, big data is relevant to using the largest possible sample to find the preferable portfolio. For instance, big data can help obtain the median, mode, and average prices to choose the best quantile for establishing the threshold (the above-mentioned comfort top and bottom, see the expression (19) which proposes the AT strategy). Additionally, big data can also find the highest or lowest values of the proposed BII indicator from a wide sample that, thanks to the law of large numbers, ensures a better representation of the results.

Regarding the relevance of big data to the development of the algorithmic procedure presented in this chapter, the chapter considers

Table 10 Stock exchanges with the lowest mobile-ratio (MR) in the sample

Stock exchanges with lowest MR	*MR*	*IUP*	*BII*
Z 1:IND	0.0167	0.02105	0.5082
VE1:IND	0.0218	0.244898	0.5107
TP1:IND	0.0526	0.244898	0.5249
QC1:IND	0.0645	0.244898	0.5302

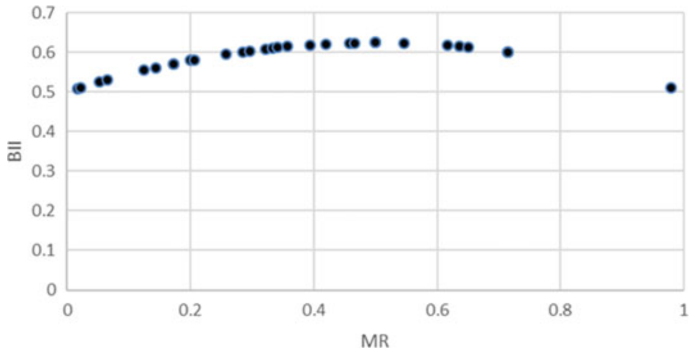


Fig. 4 Relationship between the mobile-ratio (MR) and balanced investment indicator (BII) for the sample of worldwide stock exchanges

an example using daily prices from worldwide stock exchanges per Bloomberg[1] in Table 10. This example shows that, even if the indicator is valid in most cases without using big data, it is beneficial to employ big data. For instance, when using a small sample, a Russian stock exchange (listed as VE1) appears as one of the lowest risk exchanges, as presented in Table 10, which is questionable because this country is currently suffering a period of high economic instability and uncertainty mainly due to the conflict with Ukraine. If a large sample of data were used (big data), it would help ensure more stable and trustworthy results.

Furthermore, Fig. 4 shows the relationship between the MR and BII indicators, suggesting a balance in the trade-off between profitability and low risk.

[1] Data obtained January 13, 2022.

In sum, there is incipient potential for the use of the BII algorithm presented in concert with big data, which empowers positive properties associated with the algorithm for providing an alpha in the market using the law of large numbers associated with large samples.

7 Conclusion

In conclusion, algorithmic trading (AT) is widely used by bank practitioners in making decisions associated with buying or selling financial portfolios or stocks, primarily based on tools such as big data and artificial intelligence. AT helps businesses increase profitability, provide liquidity, and reduce volatility (Virgilio, 2019). Nonetheless, there are also some risks associated with AT, such as the reduction of profits in the market when traders highly use AT, or the reduction of the perception of stock prices by investment corporates (Yan, 2019). A trading algorithm is needed that balances profitability and risks. After applying the proposed AT strategy based on the BII indicator to the prices of stocks, stock exchanges, and country indices, this chapter's algorithm offers a potential solution to this problem.

This chapter compares the results with the common use of financial trading in banks and other AT methods, hinting to a good performance of the indicator, which coincides in some cases with the investing recommendations of relevant institutions, such as the Bank of America. This proposed method is designed to present a balanced trade-off between safety and profitability regarding the correlation between the returns of a portfolio and the risk of bankruptcy by the issuer. This disjunctive is prejudicial for the investor because they have to choose between the non-default safety of a share and the received income. This dilemma is difficult to solve, but as this chapter suggests, it can be potentially navigated through the use of big data and the development of new statistical analysis techniques. The use of these novel data and techniques empowers the proposed indicators in this chapter, which weigh two modified indicators of profitability and risk of a portfolio. These indicators are economically checked by using a large sample of countries and trade data, becoming the basis for the contribution of the proposed AT strategy. After the author performs this analysis, an additional example with a smaller sample is used, which shows the importance of using a large sample when developing AT strategies. An important takeaway of this chapter is the importance of big data in developing AT strategies, which financial managers must consider when developing their own strategies.

References

Adrian, T., Crump, R. K., & Vogt, E. (2019). Nonlinearity and flight-to-safety in the risk-return trade-off for stocks and bonds. *The Journal of Finance, 74*(4), 1931–1973. https://doi.org/10.1111/jofi.12776

Aggarwal, R. K., & Samwick, A. A. (1999). The other side of the trade-off: The impact of risk on executive compensation. *Journal of Political Economy, 107*(1), 65–105. https://doi.org/10.1086/250051

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

Bloomberg. (2022, January 13). https://www.bloomberg.com/markets/stocks/futures

BME. (2022, January 14). https://www.bolsasymercados.es/esp/Estudios-Publicaciones/Estadisticas

Dunis, C. L., Laws, J., & Naïm, P. (2004). *Applied quantitative methods for trading and investment*. Wiley.

Expansión. (2022, January 13). https://www.expansion.com/ahorro/2022/01/13/61dee6d0468aebc4578b4670.html

Expansión. (2022, January 14). https://www.expansion.com/mercados/2022/01/14/61e12c8ee5fdea4a128b4661.html

Feuerriegel, S., & Prendinger, H. (2016). News-based trading strategies. *Decision Support Systems, 90*, 65–74. https://doi.org/10.1016/j.dss.2016.06.020

Hafezi, R., Shahrabi, J., & Hadavandi, E. (2015). A Bat-Neural Network Multi-agent System (BNNMAS) for stock price prediction: Case study of DAX stock price. *Applied Soft Computing, 29*, 196–210. https://doi.org/10.1016/j.asoc.2014.12.028

Hansen, K. B. (2020). The virtue of simplicity: On machine learning models in algorithmic trading. *Big Data & Society, 7*(1), 1–14. https://doi.org/10.1177/2053951720926558

Hilbert, M., & Darmon, D. (2020). How complexity and uncertainty grew with algorithmic trading. *Entropy, 22*(5), 499. https://doi.org/10.3390/e22050499

Ho, T., & Saunders, A. (1981). The determinants of bank interest margins: Theory and practice. *Journal of Financial and Quantitative Analysis, 16*, 581–600. https://www.jstor.org/stable/2330377

Huang, B., Huan, Y., Xu, L. D., Zheng, L., & Zou, Z. (2019). Automated trading systems statistical and machine learning methods and hardware implementation: a survey. *Enterprise Information Systems*, *13*(1), 132–144. https://doi.org/10.1080/17517575.2018.1493145

Hu, Y., Liu, K., Zhang, X., Su, L., Ngai, E. W. T., & Liu, M. (2015). Application of evolutionary computation for rule discovery in stock algorithmic trading:

A literature review. *Applied Soft Computing, 36*, 534–551. https://doi.org/10.1016/j.asoc.2015.07.008

Kaufman, P. J. (2005). *The new trading systems and methods*. Wiley Trading.

López-Laborda, J., & Peña, G. (2018). A new method for applying VAT to financial services. *National Tax Journal, 71*(1), 155–182. https://doi.org/10.17310/ntj.2018.1.05

Peña, G. (2020). A new trading algorithm with financial applications. *Quantitative Finance and Economics, 4*(4), 596–607. https://doi.org/10.3934/QFE.2020027

Peña, G. (2021). The key role of quoted spreads in financial services and transactions. *Economics and Business Letters, 10*(3), 208–216. https://doi.org/10.17811/ebl.10.3.2021.208-216

Pricope, T. V. (2021). *Deep reinforcement learning in quantitative algorithmic trading: A review*. arXiv preprint arXiv:2106.00123

Treleaven, P., Galas, M., & Lalchand, V. (2013). Algorithmic trading review. *Communications of the ACM*, *56*(11), 76–85.

Virgilio, G. P. M. (2019). High-frequency trading: A literature review. *Financial Markets and Portfolio Management, 33*(2), 183–208. https://doi.org/10.1007/s11408-019-00331-6

World Bank Database. (2021, June 25). https://data.worldbank.org/indicator

Yan, H. (2019). *The real effects of algorithmic trading* (Doctoral dissertation). Duke University.

High-Frequency Trading and Market Efficiency in the Moroccan Stock Market

El Mehdi Ferrouhi and Ibrahim Bouabdallaoui

1 Introduction

There is a growing interest in high-frequency trading (HFT) due to technological developments and the increasing use of technologies in financial markets. Felker et al. (2014) define HFT, a subset of algorithmic trading (Muthuswamy et al., 2010), as any consistent trading activity with a significantly brief time span, and a high number of daily discrete round turns (completed trades) and messages. Zhang (2010) describes HFT as fully automated trading strategies with very high trading volume and extremely short holding periods ranging from milliseconds to minutes and possibly hours. The United States Securities Exchange Commission (2014) defines five characteristics of HFT:

E. M. Ferrouhi (✉)
Faculty of Economics and Management, Ibn Tofail University, Kénitra, Morocco
e-mail: elmehdiferrouhi@gmail.com

I. Bouabdallaoui
Ecole Supérieure de Technologie, Mohammed V University in Rabat, Rabat, Morocco

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_4

1. Use of extraordinarily high speed and sophisticated programs for generating, routing, and executing orders;
2. Use of co-location services and individual data feeds offered by exchanges and others to minimize network and other latencies;
3. Very short time-frames for establishing and liquidating positions;
4. Submission of numerous orders that are canceled shortly after submission and;
5. Ending the trading day in as close to a flat position as possible (that is, not carrying significant, unhedged positions overnight).

Companies today are spending tens of millions of dollars on extremely large and/or complex datasets ("big data") as well as HFT infrastructure and technology, because "[m]illiseconds are millions" (Fang & Zhang, 2016). Using big data is a popular choice in today's financial community as it allows traders to make better investment choices and receive continuous feedback on their actions. The reasoning is that big data entails using large data samples, which increases confidence when making decisions. Despite the inherent volatility in financial markets, big data allows investors to quickly get a picture of the actual situation. Previously, online investing in the stock market involved making decisions based solely on market trends and calculating known risks. Today, these calculations are executed more efficiently by a computer program that can come to smarter conclusions with more information. Therefore, big data plays a primary role in decision-making for online trading, and, accordingly, we have seen an increase in resources allocated to big data in recent years.

HFT developed thanks to big data, which has, in turn, made it possible to place several hundred buy and sell orders in seconds, or even milliseconds, while optimizing risks. The development of HFT allowed investors to buy and resell stocks in a short enough period of time to avoid potentially negative market movement during the transaction.

In this study, we examine the impact of HFT on the market efficiency of the Moroccan Stock Exchange. Market efficiency implies that stock market variations are unpredictable as stock prices reflect the available information (Fama, 1995). This study, the first to use HFT in Africa, relates to one of the most developed African stock markets. Indeed, the Moroccan stock market, also known as the Casablanca Stock Exchange or the CSE, is the third oldest stock market in Africa and the third-largest by market capitalization. The small number of studies on HFT is due to the unavailability of data. This chapter offers two important contributions. First, this chapter builds on existing literature on HFT by analyzing a developing African exchange. Second, this chapter contributes to the analysis of the market efficiency of HFT. Previous studies on the Moroccan Stock Exchange, using daily data, rejected the random walk hypothesis and revealed evidence of behavioral biases such as herding behavior (Ferrouhi, 2021) and calendar anomalies (Ferrouhi et al., 2021).

Moreover, our results have policy implications concerning the accessibility to HFT. Privileged traders that have access to this data can, in theory, beat the market if markets are indeed inefficient. In the next section, we present a brief literature review of existing studies conducted on HFT and market efficiency. Next, in Sect. 3, we present our data and methodology, while in Sect. 4, we present our results. Section 5 serves as a conclusion to the chapter.

2 Literature Review

Several authors find that HFT played a positive role in improving market efficiency (Carrion, 2013; Manahov & Hudson, 2014; Martinez & Rosu, 2013), decreasing volatility (Brogaard, 2011; Hagströmer & Nordèn, 2013; Hansbrouck & Saar, 2013), and contributing to market stability (Hendershott & Riodran, 2011). However, other studies reveal the negative role of HFT in market volatility (Foucault et al., 2016; Leone & Kwabi, 2019; Zhang, 2010), price instability (Leone & Kwabi, 2019), and market efficiency (Jones, 2013). HFT's lack of availability disadvantages several market actors (Brogaard, 2011; Jones, 2013). For instance, the Australian Securities and Investments Commission (ASIC) report published in 2013 demonstrates that HFT negatively impacts the market.

The extant literature contains few studies that examine the role of HFT in price discovery and price efficiency. Thus, Brogaard et al. (2014) use data at millisecond frequencies from the NASDAQ and the New York Stock Exchange and find that HFT predicts price changes over horizons of less than three to four seconds. Meanwhile, Leone and Kwabi (2019) investigate the weak efficiency form in the FTSE100 and reject the hypothesis of a random walk at both millisecond and second frequencies, while the authors accept this hypothesis at 10, 15, 30, 60, 120, and 240-minute frequencies. Other studies that have investigated the market efficiency of HFT are concerned with stock indices in the United States (e.g. Carrion, 2013; Shafi et al., 2019; Zhang, 2010), in the United Kingdom (e.g. Leone & Kwabi, 2019); in Europe (e.g. Ammar &

Hellara, 2021; Haferkom, 2017); currencies (e.g. Manahov et al., 2014), and cryptocurrencies (e.g. Aslan & Sensoy, 2020; Sensoy, 2019), among other things. Virgilio (2019) presents a literature review on the effect of HFT on volatility, transaction costs, liquidity, price discovery, and flash crashes. This chapter will test the market efficiency in HFT and thus fill a gap in the literature.

3 Methodology and Data

This chapter aims to study the impact of high-frequency trading on market efficiency in the Moroccan Stock Exchange. The existence of market efficiency suggests that investors cannot beat the market, while a rejection of market efficiency implies that investors may be able to realize returns higher than the market by using HFT. Fama (1970) defines three forms of market efficiency:

(a) **Weak form**: prices include all the historical data,
(b) **Semi-strong form**: prices include all publicly available data, and
(c) **Strong form**: prices include both the public and private data.

The hypothesis we test in this paper is that HFT improves the market efficiency of the Moroccan Stock Exchange (at the millisecond, 1-second, 30-second, 1-minute, 2-minute, 5-minute, 10-minute, and 15-minute frequencies).

3.1 Methodology

The variance ratio test (Lo & MacKinlay, 1988) and Chow and Denning joint test (1993) are the tests commonly used to test the weak form of efficiency. The variance ratio test tests if a time series follows a random walk. The Lo and MacKinlay variance ratio $VR(k)$ is calculated as follows:

$$VR(k) = \frac{\sigma^2(k)}{\sigma^2(1)}$$

where $\sigma^2(k)$ is $1/k$ times the variance of $(X_t - X_{t-1})$, calculated as follows:

$$\sigma^2(k) = \frac{1}{m}\sum_{t=k}^{nk}(X_t - X_{t-k} - k\widehat{\mu})^2$$

and $\sigma^2(1)$ is the variance of $(X_t - X_{t-1})$, calculated as follows:

$$\sigma^2(1) = \frac{1}{nk-1}\sum_{t=1}^{nk}(X_t - X_{t-k} - \widehat{\mu})^2$$

where $m = q(nq - q + 1)(1 - kn^{-1})$
and $\widehat{\mu} = \frac{1}{nk}\sum_{t=1}^{nk}(X_t - X_{t-1}) = \frac{1}{nk}(X_{nk} - X_0)$
and $X_t = ln P_t$

P_t is the index price at t. If ratios are equal to one, then that series follows a random walk, while positive or negative ratios are synonymous with positive and negative autocorrelation. Low and Mackinlay (1988) suggest two test statistics:

- under the null hypothesis of homoskedastic increments random walk:

$$M_1(k) = \frac{VR(k) - 1}{\phi(k)^{\frac{1}{2}}}, N(0, 1)$$

where

$$\phi(k) = \frac{2(2k-1)(k-1)}{3kn}$$

- under the null hypothesis of heteroskedastic increments random walk:

$$M_2(k) = \frac{VR(k) - 1}{\phi^*(k)^{\frac{1}{2}}}, N(0, 1)$$

where

$$\phi^*(k) = \sum_{j=1}^{k-1}\left[\frac{2(k-j)}{k}\right]^2\delta(j)$$

and

$$\delta(j) = \frac{\sum_{t=j+1}^{nk} (X_t - X_{t-1} - \widehat{\mu})^2 (X_{t-j} - X_{t-j-1} - \widehat{\mu})^2}{\left[\sum_{t=1}^{nk} (X_t - X_{t-1} - \widehat{\mu})^2\right]^2}$$

Chow and Denning (1993) extend Lo and MacKinlay's variance ratio test and develop the Chow and Denning joint test. Considering a set of variance ratio estimates $\{VR(q_i)|I = 1, 2, \ldots., m\}$, where m corresponds to a set of a pre-defined number of lags and $M_1(k)$ and $M_2(k)$ presented above. The Chow and Denning joint test tests a set of sub-hypothesis:

$$H_{0i} \quad VR(k_i) = 1, \text{ for } i = 1, 2, \ldots., m$$

$$H_{1i} \quad VR(k_i) \neq 1, \text{ for } i = 1, 2, \ldots., m$$

The largest absolute value of the test statistics are

$$MV_1 = \max|M_1(k), \ 1 \leq i \leq m$$

and

$$MV_2 = \max|M_2(k), \ 1 \leq i \leq m$$

Chow and Denning (1993) follow SMM distribution (studentized maximum modulus distribution). The null hypothesis is rejected at α level of significance if MV is greater than $[1 - \left(\frac{\alpha^*}{2}\right)]$ where $\alpha^* = 1 - (1 - \alpha)^{\frac{1}{m}}$.

Following Belaire-Franch and Opong (2005) and Leone and Kwabi (2019), we use 2, 5, 10, and 30 differences, and employ the variance ratio test and Chow and Denning multiple variance ratio test using heteroskedastic robust versions of both.

3.2 Data

This chapter uses data from the MASI (Moroccan All Shares Index), the Casablanca Stock Exchange's main index, to conduct its study. As of October 7, 2021, there were 74 companies listed in the MASI, and the CSE had the third-largest market capitalization in Africa (after those of South Africa and Namibia), valued at more than USD 56 billion. We obtained data at the precision level of milliseconds under a non-disclosure agreement from the Casablanca Stock Exchange, which covers the period from August 1, 2016 (the beginning of high-frequency trading in the Moroccan stock market) to July 31, 2021. We then selected data at 1- and 30-second frequencies and at the 1-, 5-, 10-, and 15-minute frequencies. The multitude of this data allows for better knowledge of the Moroccan market in terms of risks and opportunities thanks to the shelling of the raw database, which generates the big data in various timestamps.

4 Results

According to Table 1, the skewness reveals that variations tend to be more negative for all frequencies. These results indicate that investors can realize small gains but are also exposed to big losses. The low kurtosis, observed at all time frequencies, reveals that the data exhibits less extreme values. Additionally, the results of the Jarque–Bera test allow us to reject the normality hypothesis, which indicates that the distribution of values does not follow a Gaussian distribution.

As presented in Sect. 3, variance ratios above one indicate that markets are inefficient, while ratios equal to one indicate weak market efficiency. As Table 2 illustrates below, the Lo and MacKinlay variance ratio exceeds one at the millisecond, 1-second, and 30-second time frequencies, and so we reject the random walk hypothesis and thus reject the hypothesis of market efficiency for these frequencies. The results of the Chow and Denning joint test further confirm this outcome. Thus, we conclude that privileged investors with access to HFT can predict future prices, and so the possibility of arbitrage exists in this market at these time frequencies. Our results are thus in alignment with those of Brogaard et al. (2014) and Leone and Kwabi (2019).

However, results at 1-, 2-, 5-, 10-, and 15-minute frequencies show evidence of a random walk. As a result, we conclude that HFT improves the market efficiency of the Moroccan stock exchange at low time frequencies (1-, 2-, 5-, 10-, and 15-minute frequencies). These results confirm those obtained in developed markets (Carrion, 2013; Manahov & Hudson, 2014; Martinez & Rosu, 2013).

Nevertheless, as the access to high-frequency trading (millisecond, 1-second, and 30-second frequencies) is limited, only investors having access to such data can gain an advantage over the market by predicting future variations. Thus, we remark that the market tends to be more efficient as the data frequency decreases. These findings confirm previous

Table 1 Descriptive statistics of Moroccan All Shares Index (MASI) price tick changes by different time-frequencies

Statistic	*Millisecond*	*1 second*	*30 seconds*	*1 minute*	*5 minutes*	*10 minutes*	*15 minutes*
Mean	11,610.17	11,604.43	11,576.05	11,557.96	11,518.91	11,511.11	11,507.83
Median	11,668.82	11,663.62	11,605.09	11,575.13	11,536.76	11,529.56	11,527.77
Max	13,387.50	13,387.50	13,384.63	13,384.63	13,384.63	13,382.29	13,384.63
Min	8916.39	8916.39	8916.39	8916.39	8920.35	8939.13	8939.13
Std. Dev	920.47	921.32	917.52	913.82	899.00	895.02	895.22
Skewness	−0.52	−0.52	−0.46	-−0.42	−0.34	−0.33	−0.33
Kurtosis	2.66	2.66	2.62	2.61	2.61	2.63	2.63
Jarque–Bera	26,535.57	23,473.20	12,550.08	8404.78	2071.75	1065.39	749.57
Probability	0	0	0	0	0	0	0
Observations	532,940	474,491	305,878	233,189	81,394	44,676	31,774

Table 2 Variance ratio test results

	Millisecond				**1 second**			
Joint tests		**Value**	**df**	**Prob.**		**Value**	**df**	**Prob.**
		6.009337	532,939	0.0000		5.503556	474,490	0.0000
Individual Tests					*Individual Tests*			
Period	*Var. ratio*	*Std. Error*	*z-statistic*	*Prob.*	*Var. ratio*	*Std. Error*	*z-statistic*	*Prob.*
2	1.006605	0.001975	3.344341	0.0008	1.005066	0.002096	2.417239	0.0156
5	1.023103	0.004948	4.668876	0.0000	1.019712	0.005223	3.773813	0.0002
10	1.051018	0.008704	5.861777	0.0000	1.047540	0.009135	5.204486	0.0000
30	1.094186	0.015673	6.009337	0.0000	1.089892	0.016333	5.503556	0.0000
		30 seconds				**1 minute**		
Joint tests		**Value**	**df**	**Prob.**		**Value**	**df**	**Prob.**
		4.611472	305,877	0.0000		0.884236	233,188	0.8489
Individual Tests						*Individual Tests*		
Period	*Var. ratio*	*Std. Error*	*z-statistic*	*Prob.*	*Var. ratio*	*Std. Error*	*z-statistic*	*Prob.*
2	1.043399	0.009411	4.611472	0.0000	0.997093	0.004181	−0.695377	0.4868
5	1.061528	0.015858	3.879851	0.0001	0.995018	0.008526	−0.584344	0.5590
10	1.063017	0.019267	3.270672	0.0011	0.989388	0.012001	−0.884236	0.3766
30	1.042480	0.024842	1.709997	0.0873	0.981024	0.024316	−0.780414	0.4351
		2 minutes				**5 minutes**		
Joint tests		**Value**	**df**	**Prob.**		**Value**	**df**	**Prob.**
		1.082669	159,382	0.7297		3.269614	81,393	0.4300
Individual Tests						*Individual Tests*		
Period	*Var. ratio*	*Std. Error*	*z-statistic*	*Prob.*	*Var. ratio*	*Std. Error*	*z-statistic*	*Prob.*
2	0.999369	0.004585	−0.137591	0.8906	1.000720	0.006407	0.112326	0.9106
5	0.997895	0.009662	−0.217850	0.8275	0.997669	0.015224	−0.153085	0.8783
10	0.993289	0.015708	−0.427225	0.6692	1.024301	0.026749	0.908457	0.3636
30	1.035162	0.032477	1.082669	0.2790	1.144573	0.044217	3.269614	0.2311

(continued)

Table 2 (continued)

		10 minutes				**15 minutes**		
Joint tests		**Value**	**df**	**Prob.**		**Value**	**df**	**Prob.**
		4.504799	44,675	0.2150		5.587624	31,773	0. 1660
Individual Tests						*Individual Tests*		
Period	*Var. ratio*	*Std. Error*	*z-statistic*	*Prob.*	*Var. ratio*	*Std. Error*	*z-statistic*	*Prob.*
2	0.991418	0.009241	−0.928748	0.3530	1.022375	0.015750	1.420666	0.2854
5	1.026134	0.026086	1.001834	0.3164	1.066856	0.030802	2.170502	0.2500
10	1.088575	0.038896	2.277231	0.2828	1.138655	0.042770	3.241892	0.2082
30	1.257455	0.057151	4.504799	0.1929	1.353973	0.063350	5.587624	0.1500

results that the unavailability of HFT disadvantages traders who do not have access to data (Brogaard, 2011; Jones, 2013; Leone & Kwabi, 2019). Thus, the stock market authority for this market (the Moroccan Capital Market Authority) should implement and strengthen measures to protect investors by ensuring that all the information is freely and equally available to investors.

5 Conclusion

This chapter studied the impact of high-frequency trading on market efficiency in the Moroccan Stock Exchange. We used data covering five years, from 2016 to 2021, with frequencies between a millisecond and 15 minutes. We find no evidence of a random walk at the millisecond, 1-second, nor 30-second frequencies. Therefore, it follows that investors with a faster market connection and an efficient algorithm can use privileged information to realize returns higher than those of the market. However, we find evidence of market efficiency at 1, 2, 5, 10, and 15-minute time frequencies. We thus conclude that there is a possibility of arbitrage for investors in this market at higher temporal frequencies. Our main recommendations are that the Moroccan Capital Market Authority implement and strengthen measures to protect investors and increase available information to improve market efficiency, especially at 1-millisecond, 1-second, and 30-second frequencies.

One area in the future to look out for is the growth of big data in financial markets, as it will be useful to analyze our study's data in such a way that time is understood as an important influence on price variations. Such inferences from big data could introduce the application of forecasting and machine learning tools to create accurate models that can be applied and used to forecast future variations more effectively than today. The growth of big data will allow for real-time data analysis from multiple sources (e.g., financial brokers) before executing a trade. Moreover, correlating real-time data with historical data sources would help identify more profitable trades with a higher degree of accuracy; real-time data processing technologies, such as in-memory data grids and processing engines, could assist in this identification, so long as the analysis is made at high frequencies. Investors who can best integrate HFT with big data technologies at high temporal frequencies will be most likely to gain an edge in the financial marketplace in the future.

Acknowledgements Many thanks to Ayoub Youssefi, Head of the Development Department in the Casablanca Stock Exchange, for sharing the data used in this study. We also thank our colleague Rachida El Guerraoui for proofreading and our reviewers for their valuable comments.

References

Ammar, I. B., & Hellara, S. (2021). Intraday interactions between high-frequency trading and price efficiency. *Finance Research Letters, 41*, 101862.

Aslan, A., & Sensoy, A. (2020). Intraday efficiency-frequency nexus in the cryptocurrency markets. *Finance Research Letters, 35*, 101298.

Australian Securities and Investments Commission. (2013). *Dark liquidity and high-frequency trading*.

Belaire-Franch, J., & Opong, K. K. (2005). A variance ratio test of the behaviour of some FTSE equity indices using ranks and signs. *Review of Quantitative Finance and Accounting, 24*(1), 93–107.

Brogaard, J. (2011). *High frequency trading and volatility*. SSRN eLibrary.

Brogaard, J., Hendershott, T., & Riordan, R. (2014). High-frequency trading and price discovery. *The Review of Financial Studies, 27*(8), 2267–2306.

Carrion, A. (2013). Very fast money: High-frequency trading on NASDAQ. *Journal of Financial Markets, 16*, 680–711.

Chow, K. V., & Denning, K. C. (1993). A simple multiple variance ratio test. *Journal of Econometrics, 58*(3), 385–401.

Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. *The Journal of Finance, 25*(2), 383–417.

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

Fang, B., & Zhang, P. (2016). Big data in finance. In *Big data concepts, theories, and applications* (pp. 391–412). Springer.

Felker, T., Mazalov, V., & Watt, S. M. (2014). Distance-based high-frequency trading. *Procedia Computer Science, 29*, 2055–2064.

Ferrouhi, E. M. (2021). Herding behavior in the Moroccan stock exchange. *Journal of African Business, 22*(3), 309–319.

Ferrouhi, E. M., Kharbouch, O., Aguenaou, S., & Naeem, M. (2021). Calendar anomalies in African stock markets. *Cogent Economics & Finance, 9*(1), 1978639.

Foucault, T., Hombert, J., & Roşu, I. (2016). News trading and speed. *The Journal of Finance, 71*(1), 335–382.

Haferkom, M. (2017). High-frequency trading and its role in fragmented markets. *Journal of Information Technology, 32*(3), 283–296.

Hagströmer, B., & Nordèn, L. (2013). The diversity of high frequency traders. *Journal of Financial Markets, 16*, 741–770.

Hansbrouck, J., & Saar, G. (2013). Low-latency trading. *Journal of Financial Markets, 16*, 646–679.

Hendershott, T., & Riordan, R. (2011). *High frequency trading and price discovery* (p. 3). University of California.

Jones, C. M. (2013). *What do we know about high-frequency trading?* (Columbia Business School Research Paper, 13-11).

Leone, V., & Kwabi, F. (2019). High frequency trading, price discovery and market efficiency in the FTSE100. *Economics Letters, 181*, 174–177.

Lo, A. W., & MacKinlay, A. C. (1988). Stock market prices do not follow random walks: Evidence from a simple specification test. *The Review of Financial Studies, 1*(1), 41–66.

Manahov, V., & Hudson, R. (2014). The implications of high-frequency trading on market efficiency and price discovery. *Applied Economics Letters, 21*(16), 1148–1151.

Manahov, V., Hudson, R., & Gebka, B. (2014). Does high frequency trading affect technical analysis and market efficiency? And if so, how? *Journal of International Financial Markets, Institutions and Money, 28*, 131–157.

Martinez, V. H., & Rosu, I. (2013, January). *High frequency traders, news and volatility* (AFA 2013 San Diego Meetings Paper).

Muthuswamy, J., Palmer, J., Richie, N., & Webb, R. (2010). High-frequency trading: Implications for markets, regulators, and efficiency. *The Journal of Trading, 6*(1), 87–97.

Sensoy, A. (2019). The inefficiency of Bitcoin revisited: A high-frequency analysis with alternative currencies. *Finance Research Letters, 28*, 68–73.

Shafi, K., Latif, N., Shad, S. A., & Idrees, Z. (2019). High-frequency trading: Inverse relationship of the financial markets. *Physica A: Statistical Mechanics and Its Applications, 527*, 121067.

US Securities and Exchange Commission. (2014). *Equity market structure literature review, part II: High frequency trading*.

Virgilio, G. P. M. (2019). High-frequency trading: A literature review. *Financial Markets and Portfolio Management, 33*(2), 183–208.

Zhang, F. (2010). *High-frequency trading, stock volatility, and price discovery*. SSRN 1691679.

Ensemble Models Using Symbolic Regression and Genetic Programming for Uncertainty Estimation in ESG and Alternative Investments

Percy Venegas, Isabel Britez, and Fernand Gobet

1 Introduction

Navigating bright spaces is much easier than walking in dark and foggy alleys. Recently, we have witnessed the growing clamor for clarity in the alternative investment space for sustainable strategies and environmental, social, and governance (ESG) investment.[1] Many questions are being

[1] Fallender (2020) emphasized the novelty of the ESG acronym and the need for clarity so that the ESG investment space is clearer to entrants: "I think there are still a lot of people. And think of large companies like Intel who've been working on these issues for many years. But there are a lot of companies out there to which it's relatively new. It's an acronym. So I think it's also about language and understanding how to kind of frame it for people in the language that they understand".

P. Venegas (✉)
Economy Monitor, King's College London, London, UK
e-mail: v.percy@economymonitor.com

I. Britez
FairAi, University of Cambridge, Cambridge, UK
e-mail: isabel@fairai.uk

F. Gobet
The London School of Economics and Political Science, London, UK
e-mail: f.gobet@lse.ac.uk

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_5

raised by the financial community, such as: what is an ESG strategy and investment practice, how reliable are current ESG ratings, and is ESG investment capable of producing outstanding returns? Although we do not expect to answer all of these questions in this chapter, we intend to provide clarity within the alternative investment space.

To achieve this purpose, we discuss two main topics in this chapter: the first is how companies can integrate the ESG approach and achieve their objectives through risk reduction. This ESG integration would allow us to understand how stocks and exchange-traded funds (ETFs) relate to ESG scores and the relevant information and data gaps that need to be filled by researchers. Second, we use a symbolic regression (SR) approach to address the modeling uncertainties of the ESG data currently available for the same asset classes.

As we intend to discuss the alternative investment space, the stocks to be modeled are sustainable ETFs and publicly traded private equity (PE) stocks. Sustainable ETFs are intended by their creators to provide exposure to ESG-oriented stocks. If the stocks that are part of the ETF have uncertain ESG scores, that uncertainty percolates up into the ETF itself, and so it is essential to understand the risk to investors. Finally, we present private equity (PE) as an alternative investment asset class offering exposure to ESGs and increased money flows.

In the second part of the chapter, for each asset class introduced, we analyze symbolic regression models that present latent uncertainty to develop a trustable risk analysis for alternative investments. We also offer some suggestions of variables that are presently more relevant to cash flow. The models developed in this part of the chapter elucidate the alternative market structure and investor preferences in sustainable finance.

As ensemble models have emerged in recent years as a viable way to understand prediction uncertainty, practitioners in computational finance seeking to gain expertise in the field will find the methodologies described helpful for application in their work. Symbolic regression via genetic programming is an alternative approach to those ranging from statistical learning to other forms of evolutionary computation.

Sand (2021) raises the issue of ESG and returns: "Many believe it [ESG investment] does come at the expense of returns; in the US many sponsors don't take the topic seriously, but the data will suggest otherwise and we've started to try to build something to evaluate this".

In the next section, we present a background on stocks and ETFs. In Sects. 3 and 4, we present the modeling methodology and results, respectively, using an SR approach. Section 5 provides a discussion of the results and their implications. Lastly, we conclude the chapter in Sect. 6.

2 Background

2.1 Stocks and ETFs

Factors that drive investment decisions are the expected return on investment and the investor's risk tolerance. ESG adds a third layer to the investment decision-making process—considering environmental, social, and governance factors alongside financial factors (Briand, 2022). The ESG concept was launched by the United Nations in 2005 (The Global Compact, 2005). More recently, there has been a surge of interest in the topic from companies, investors, and asset managers, as seen in Fig. 1.[2] Expectations of sustainable financial practices from the public have become higher since the 2021 COP26 summit, along with a commitment to a sustainable world across many sectors (Thorne, 2021).

ESG is also an additional risk factor for companies and investors. As examples, we explore two public companies, addressing first the environmental theme (Levi Strauss) and then the social and governmental theme (British American Tobacco). In evaluating these examples, we raise the relevance of using the ESG framework to anticipate risks and reduce business uncertainties.

2.1.1 Levi Strauss

First, suppose we invest in a clothing company like Levi Strauss (NYSE: LEVI). According to fashion activism watchdog Fashion Revolution (Wheeler, 2019), a single pair of Levi's 501 jeans requires 3,781 L of water for its production, exacerbating the already severe issue of water scarcity. How can Levi Strauss manage production sustainably?

Agudo (2022) raises the question of current data gaps in the market: "There is a clear demand from the market to move to more sustainable data, but we are lacking accuracy, transparency, and data and this is where we need to move to have a clearer position."

[2] In 2019, more than 630 investors collectively managing more than $37 trillion signed the Global Investor Statement to Governments on Climate Change urging governments to require climate-related financial reporting (GIS, 2019).

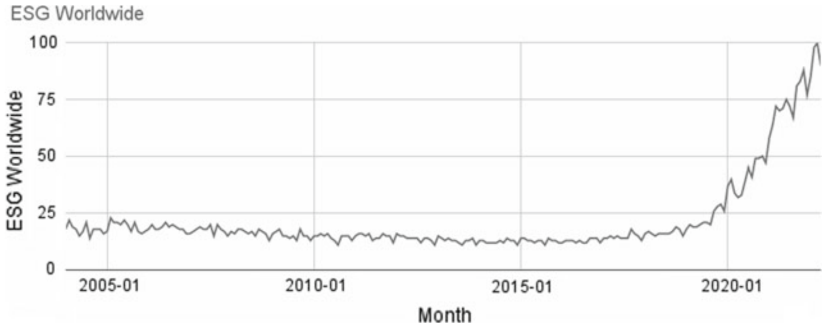


Fig. 1 'ESG' (environmental, social, and corporate governance) searches by month (*Source* Google Trends, 2022. https://trends.google.pt/trends/explore?date=all&q=esg)

Some actions can be taken to integrate long-term growth focused on efficiency into corporate strategy, thereby securing a future for a company. For example, new technologies could be developed to reduce the water required for manufacturing. Water could also be recycled. Alternatively, the business model could shift, focusing instead on cotton-made pants requiring fewer resources.

For ESG integration, identifying a material issue[3] should be the first step for a company. In the Levi's example, the material issue was water scarcity which would affect their entire production. Eccles and Serafeim (2013) argue that a company should understand the material factors that may destabilize shareholder value to ensure that it can continue to deliver that value during any crisis. We suggest two main pillars for this analysis: the company sector and its strategy.[4] For example, it is harder for a

[3] According to the US Securities and Exchange Commission (2022, p. 7), information is material if there is a substantial likelihood that a reasonable investor would consider it important in deciding how to vote or make. From the SEC 2010 docket, it is understood that "these [material] effects can impact a registrant's personnel, physical assets, supply chain and distribution chain." They can include the impact of changes in weather patterns, such as increases in storm intensity, sea-level rise, melting of permafrost and temperature extremes on facilities or operations. Changes in the availability or quality of water, or other natural resources on which the registrant's business depends, or damage to facilities or decreased efficiency of equipment can have material effects on companies.

[4] According to the 2022s (2022, p. 290) resolution, from 2024, it will be mandatory for public companies to disclose their company data on climate change risk. In 2021,

coal-fired utility to flourish under low permitted greenhouse gas (GHG) emissions than for a bank. Similarly, a company that uses highly-skilled workers would be less likely to have to deal with wage demands than one that is dependent on low-cost labor (strategy).

2.1.2 British American Tobacco (BATS)

British American Tobacco (LON: BATS), the largest tobacco company in the world, has been publicly traded since 1989. It reported increasing returns until reaching its peak on May 26th, 2017. Since then, its stocks have lost 51.1% of their value, underperforming the FTSE 100 during the same period by 49 percentage points. Two significant factors account for these numbers. First, the steep decline in 2017 correlates with a controversy involving BATS governance. In 2017, BATS was the subject of a fraud investigation that documented corruption in West Africa, involving illegal payments to politicians, the police, and even workers from a competitor (Awasthi & Muvija, 2017).

Second, BATS's poor performance is associated with social demand for healthier alternatives to smoking, such as electronic cigarettes. ESG factors incorporate societal values and reflect the fact that societal changes can materially impact company success; in this respect, anticipating consumer behavior changes and being ready for them in terms of products and deliverables results through the innovation of its products and services results in more consistent company performance. The aim is to anticipate future societal demands and to be able to react appropriately to them, as was apparent in the case of BATS.

These ESG factors should translate from the strategy and actions of a company directly to investor decision-making processes, but this is currently happening mainly through the lenses of rating agencies. These agencies classify companies and industries and score them according to two criteria. The first criterion is the category weight for each ESG rating (e.g., for the retail industry, governance and social factors will carry greater weight due to their materiality) (S&P Global, 2021). The second criterion is the company's compliance with metrics that investors can use to track the ESG parameters being followed by the company.

The European Union published its first Taxonomy on ESG. We expect public companies to have a greater degree of open climate and ESG risk information. The relevance of increasing data for this research is needed to further model through SR and risk comprehension.

The way its components are expressed and measured is still impacting investors' ability to use ESG as an investment factor. There are many rating companies and systems in the market raising hundreds of ESG data points, but there is no consensus on how this rating system should be done by agencies, as there is still no unified rating system. Howard-Grenville (2021) notes that the correlations between the different ESG ratings are currently not as large as expected (from 0.5 to 0.6) and is not confident that this can improve in the future.

Conflicting ratings obscure the path from metrics to best actions.[5] Peterson and Neuville (2021) exemplify the conflicting ESG parameters with the way Tesla (TSLA) is scored by rating agencies. While the environmental score of the Financial Times Stock Exchange (FTSE) considers Tesla a zero, Morgan Stanley Capital International (MSCI) nearly grants it a 100. The reason behind this rating difference is the focus placed by different rating agencies on different aspects that go into an ESG rating. The FTSE is looking at Tesla's supply chain, and due to the lack of reported data on its emissions, the rating received is zero. On the other hand, MSCI focuses on measuring product emissions and understands them to be emission-free, leading Tesla to a higher environmental score.

India is the first government to suggest that ratings should be standardized and providers regulated based on a variety of risk factors, as Sebi (2022) states below:

> Since the activities of ESG ratings providers (ERPs) are typically not subject to regulatory oversight at present, increasing reliance on such unregulated ESG rating providers in securities markets raises concerns about the potential risks it poses to investor protection, the transparency and efficiency of markets, risk pricing, and capital allocation, among others. Moreover, a lack of transparency in this area gives rise to the risk of greenwashing and misallocation of assets which could lead to infirmity in such ESG rating and a consequent lack of trust thereof. Therefore, there arises an imperative need, more than ever before, to ensure that the providers of such

[5] Some of the main ratings available now are CSA, from S&P; GIC's sustainability; ISE from B3; Paris agreement commitment; SDGs commitment; UNGP participation and the IRIS + system; Revinitiv ESG score; Marketpsych scores; FTSE; and MSCI. One example of the kind of objections leveled at these ratings is the one that falls under the use of the SDGs commitment: it is suggested that SDGs should be developed through governmental initiatives.

products operate in a transparent and regulated environment that balances the needs of all stakeholders (Sebi, 2022, p. 2).

Besides the smaller than expected correlation between ratings and regulation of rating providers, another concern relates to the capability of measuring environmental, social, and governance factors. Currently, we lack consistency between ratings[6] because of measurement accessibility. It is also unclear to what extent social or governance metrics should be considered.

In addition to these issues, the amount of data available for these ratings is scarce. Currently, rating agencies release data updates monthly, but this still does not allow access to ESG data on every public company. Moreover, even if we did have access to all ESG data for each company, looking into their supply chain partners would provide an additional complication when factoring them into the rating. For example, Amazon has more than six hundred suppliers from forty different countries, which would make it almost impossible to gage what their proper ESG rating should be. For the purpose of this discussion, we understand that the growth and increasing relevance of this market can incentivize data improvement and expansion, which is necessary for us to expand our analytic capabilities and reduce uncertainty in the ESG rating space.[7]

2.1.3 How ETFs Integrate ESG Factors into Stock Selections

Now that we have considered the three major issues concerning the ESG rating factors—lack of unified rating system, measurement capabilities,

[6] Cheong (2022): "There is a lot of lack of consistency in environmental and social reporting, and it's an area where we have been promoting and calling for some standardization. We often look at a variety of sources, and as S&P global, we have other divisions of the company that provide valuable data. Truecost, for example, aggregates environmental data; but more importantly, we have a relation with company management, so we'll often get data from company management itself; we have the ESG scores, and we sort of triangulate all of the sources".

[7] From 2022 onward, the UK government is preparing to make mandatory "ESG reporting from all UK private companies and LLPs [Limited Liability Partnerships] with more than 500 employees, as well as all publicly quoted UK companies" (Glover, 2021). Increasingly, considering ESG factors alongside financial factors is no longer an option. As we see the United States Security Exchange Commission, the Indian Security Exchange Board, and the European Union's movement toward ESG regulation for public companies, we can expect to know more about ESG investing in the future, easing the alignment between investments and ESG criteria.

and lack of data—we can return to discussing the ETF asset class by understanding how it integrates ESG factors with stock selections. Moreover, we can also discuss the additional uncertainty present when bundles of stocks are used to produce a single ESG rating for an ETF.

Index funds and ETFs are means for “buying the market” by tracking and investing in the market index. These indexes vary, from those indexed to the country’s inflation rate to commodities fluctuations and sets of stock yields. One of the main indexes in the United States (U.S.) is the Standard and Poor’s 500 (S&P 500), an index which tracks 500 significant U.S. stocks. Sustainable ETFs follow the returns of an index of sustainable company stocks.[8]

In the U.S. (ETF Database), there are currently more than 2000 ’responsible investing’ ETFs available, classified into three main categories:

a. **Sustainable impact solutions:** which considers the percentage of revenue of the portfolio generated by goods and services related to sustainable impact solutions;
b. **ESG score of the underlying holdings:** focusing on scores instead of revenue; and
c. **Weighted average carbon intensity (WACI):** the amount of greenhouse gas emissions per revenue unit created.

The responsible investing list presents ETFs with rates ranging from high to low, making it possible to find bad performers in each of the categories listed above. The ETF selection for this list is made through exclusion, meaning that, for all U.S. ETFs, only those whose metrics which do not comply with ETF database guidelines are removed. These exclusion factors include direct involvement with predatory lending, controversial weapons, and alcohol or tobacco products.

One of the drawbacks to this list-building methodology is that in it, we find listed ETFs with poor ESG ratings, high GHG emissions, and low-profit creation from sustainable solutions. Even though an individual stock rating may be poor, when the stocks are pooled collectively in an ETF, they have crossed a minimum threshold and are thus included on

[8] “Well-run index funds track the market almost exactly and charge very low management fees, often less than 0.1% per year. (…) they have attracted about $2 trillion from investors” (Brealey et al., 2017, p. 180).

the list. Thus, excluding only the least responsible ETFs results in an inflated range of responsible investing options remaining on the list. It makes better sense to give investors à la carte information about impact-investment ETFs so that, depending on investor preferences, investors can customize their portfolio allocation.

The ESG data concerns mentioned in the previous section on stocks—standardization and access to ESG metrics—can also be incorporated when analyzing sustainable ETFs, but with the added concern of the multiplying effect of the index. If rating uncertainty for one company is an issue, then ESG ETFs aggregating hundreds of ratings (one for each stock) are even more problematic. If the individual stocks that are part of the ETF have uncertain ESG scores, that uncertainty can make its way into the ETF itself (perhaps in nonlinear ways), which needs to be carefully considered for investors to understand their real exposure.

3 Modeling and Data Collection

3.1 Modeling

This chapter uses the implementation of model ensembles using symbolic regression (SR) and genetic programming proposed by Kotanchek et al. (2008). Researchers have previously applied this method to study financial infrastructures and alternative digital assets (Venegas, 2021). Our focus here is on predictive uncertainty estimation, which means that instead of intending to predict ESG scores and similar quantities for trading purposes, we apply the technique to model these target variables to quantify the uncertainty of the estimation and thus better understand the existing market structure. Our models were created using an evolutionary search approach using symbolic regression and genetic programming. We then used an ensemble approach to create numerous diverse models using the DataModeler software package.

3.1.1 Symbolic Regression (SR)

Traditionally, using regression models entails establishing the form of the model and selecting the parameters that best suit the observed data. Alternatively, symbolic regression identifies the model form that most accurately captures the behavior of the data. Essentially, we allow the data to dictate the structure of the model rather than impose a priori assumptions. A preference for simplicity aids our model search as a good model should be accurate and devoid of superfluous structures and variables.

3.1.2 Symbolic Regression Versus Regression Models

The main benefit of using symbolic regression rather than linear regression is that we do not have to make a priori assumptions about the model form. Instead, using an evolutionary process will inherently identify the driving variables and their correlations to one another. Using this process will also automatically model the behavior of the targeted response variable. A second benefit of using symbolic regression is that it can reveal if a system is truly linear. In other words, we will have a level of insight, comprehension, and trustworthiness that is difficult to achieve with linear regression alone. In addition, the symbolic regression algorithm helps avoid human errors in selecting parameters when modeling using methods such as linear regression (Kotanchek et al., 2008).

Moreover, other types of regression models use either simplifying assumptions (e.g., about the order of the polynomials) or have the potential for human biases to impact the model form, as previously mentioned. These models then seek to optimize the parameters using these template model structures. On the other hand, symbolic regression seeks to uncover the structure and the related parameters. The absence of requirements for domain reasonableness is both a strength and a weakness, which does not always produce unexpected and illuminating results. Additionally, numerous solutions for symbolic regression may be developed by users because an infinite number of models can fit a data set. The user can select the optimal model(s) for their purposes during post-processing.

There is a fundamental difference between assuming a model form and its parameters before modeling compared to post-processing selection of the evolutionary models developed. In particular, selecting models in post-processing allows the use of particular financial domain expertise to provide context without imposing arbitrary structures a priori—the nature of the process dictates the model form. Artificial intelligence pioneer Herbert Simon states this clearly from the perspective of behavioral and cognitive Financial: "Everyone designs who devises courses of action aimed at changing existing situations into preferred ones" (Herbert A. Simon. The Sciences of the Artificial, Third Edition, 1996, pp. 111–167).

In this sense, it is preferable to apply domain knowledge to the model selection rather than dictate what form the model should appear.

3.1.3 Trustable Model Ensembles

Trustable models are ensembles (models comprised of a group of models) that feature extrapolations into regions of unknown parameter space (e.g., within the space of possible values, those that have not been observed). Since the models are developed by users using an evolutionary approach (genetic programming), ensembles can take many potential forms. They are achievable as a result of the various model structures hypothesized, explored, and refined during the model search process. When it comes to post-processing, there are a variety of models to choose from based on the fitted models.

One of the main benefits of ensemble models is that they provide a measurement of consensus while allowing for a diversity of models. The ensembles will agree (converge to similar values) if they operate within the data regions used in their development. However, they will tend to disagree if the user asks them to operate outside that data region or if the underlying system has undergone fundamental changes.

3.1.4 Estimating (and Reducing) Uncertainty

The first uncertainty that comes with estimation using ensembles comes from the ranges of the predicted target values in each computed prediction—wider spreads indicate a larger uncertainty in the predictions. However, since ensembles are developed by users using a limited dataset of what data providers supply at any given moment, and the universe of companies and investable assets grows constantly, further modeling rounds are necessary to reduce uncertainty. At this point, two steps are necessary to use these models:

1. In order to validate or refute the predicted optimal response and remove uncertainty from the model, the user must collect additional data; in other words, the information richness of each data sample obtained must be optimized by the user.
2. The user must build additional model ensembles and repeat the previous step until satisfactory results are attained. The market for ESG metrics is very active, with new and more extensive data sets made available constantly, and so we should not expect to hit a wall in terms of access to materials for model development in the foreseeable future.

3.2 Data Collection

We gathered datasets from specialized providers of private companies and exchange-traded funds and augmented those with ESG scores from sustainability ranking providers. The assets belong to the alternative investments category (other than stocks, bonds, and cash). The software used to carry out the study is Mathematica (Wolfram Research, 2021), and the genetic programming package is DataModeler (Evolved Analytics, 2021). The datasets contained numerical and categorical data, and we handled all data preparation using DataModeler.[9] For simplicity, we have provided definitions only for the variables included in the final model ensembles (see Appendix, Table 1). The dataset providers are Pitchbook, Refinitiv, and ETF DB.

3.2.1 Publicly Traded Private Equity Stocks

In our analysis, the dimensions of the dataset are 37 (companies) × 45 (variables). The target variable is the *ESG Score*,[10] provided by Refinitiv. All other private company data points come from Pitchbook. The companies are peers in the extractive industries (mining and metal-related manufacturing), which traditionally have experienced challenges regarding ESG practices.

3.2.2 Sustainable ETFs

For this modeling exercise, we obtained a list of ETFs previously classified as sustainable by ETFdb, a database with proprietary data and aggregates scores from other data providers, including ESG scores. The dataset dimensions are 5 rows (ETFs) × 48 columns (variables). The target is the *ESG Score*,[11] a numerical variable.

[9] Source Datasets and Predictions Tables are Available at https://www.autonomous.economymonitor.com/s/Evolving-Algorithms-for-Uncertainty-Estimation-in-ESG-and-Alternative-Investments-Supplementary-Mate.zip.

[10] Refinitiv ESG scores measure the ESG performance of companies, based on reported data in the public domain across three pillars and 10 different ESG topics. Refinitiv ESG combined score is an overall company score based on the reported information in the environmental, social and corporate governance pillars (ESG Score) with an ESG Controversies overlay.

[11] MSCI ESG Quality Score. This is the overall ESG score. It measures the ability of underlying holdings to manage key medium to long-term risks and opportunities arising from environmental, social, and governance factors.

4 Results

4.1 Publicly Traded Private Equity Stocks

An investment firm that makes investments by following one or more of the various private equity strategies listed on a public stock exchange is referred to as publicly traded private equity. Since systematic ESG ratings for private companies are less common than for stocks, the choice of a publicly traded private equity dataset allows us to remain in the realm of alternative investments while still using ESG ratings.

There were 11 driver variables (i.e., inputs that were present in a predefined fraction of models) discovered in the first modeling round: *Primary Business Sector, Business Status, Employees, Year Founded, Gross Profit, Net Income, Enterprise Value, EBIT, Net Debt, Number of Active Investors*, and *Last Financing Status*. The target variable, ESG scores, for a first approximation, depended on all of these factors.

In the second modeling round, we decided to optimize for linear models by avoiding exponential terms in the formula. In practice, this operation favors additive structures in mathematical expressions rather than multiplicative structures. We then remove variables that did not pass an ANOVA p-value test, allowing us to determine if the differences between the means are statistically significant. The second ensemble model thus contained a higher fraction of optimized linear models.

The variable set in the second modeling round was smaller (8 variables) than in the first round, as we obtained eight variables in the former versus 11 variables in the latter. The 11 driver variables in this round were: *Employees, Year Founded, Net Income, Enterprise Value, EBITDA, Market Cap, Net Debt*, and *Number of Active Investors*. This reduction in driver variables from the first round suggested that fewer inputs were necessary to arrive at a prediction; therefore, the ensemble model was simplified. This simplification indicates to us that if we had stayed with the mathematical expression obtained in the first round, we could have achieved similar results using more input variables. However, the cost of monitoring more inputs in real life is high, especially when working with financial data, and so this is not recommended. The Pareto plots in Fig. 2 compares the first ensemble model (ensemble) and the linear-optimized ensemble model (ensemble2). The figure indicates that the second ensemble model exhibits less complexity than the first ensemble model.

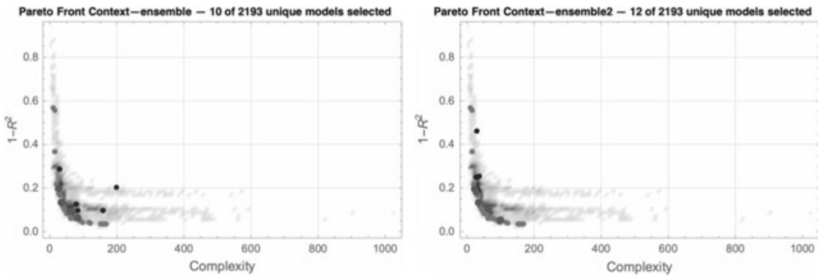


Fig. 2 Accuracy-complexity trade-off across modeling rounds (*Note* Darker Points are optimal models for Private Equity)

The prediction plots in Fig. 3 show more clearly another benefit of the less-complex model structure: ensemble 2 has no outliers (lighter-tone points) and therefore offers more certainty around points that would otherwise be difficult to model and generate a prediction for. At the same time, the second ensemble appears to be more uncertainty-aware because prediction ranges vary more widely across different response values. That is, by using a group of models (an ensemble) and a subset of variables (*Employees, Year Founded, Net Income, Enterprise Value, EBITDA, Market Cap, Net Debt*, and *Number of Active Investors)* to model *ESG Score*, we obtain a model that is aware of its accuracy in the sense that it includes predictive ranges. This characteristic of ensemble modeling is material to finance practitioners, who not only need to report on financial performance but also provide measurements for risk management.

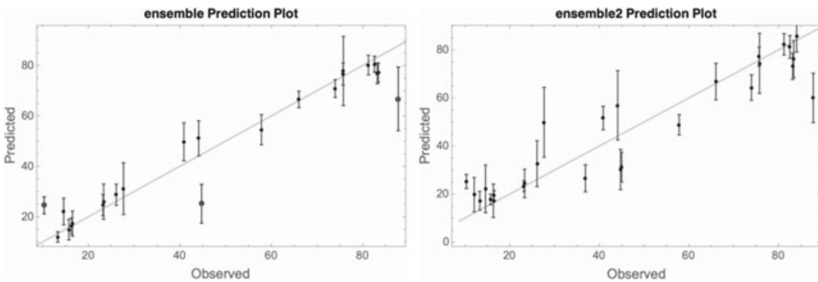


Fig. 3 Ensemble prediction plots, modelling rounds comparison for publicly traded private equity stocks

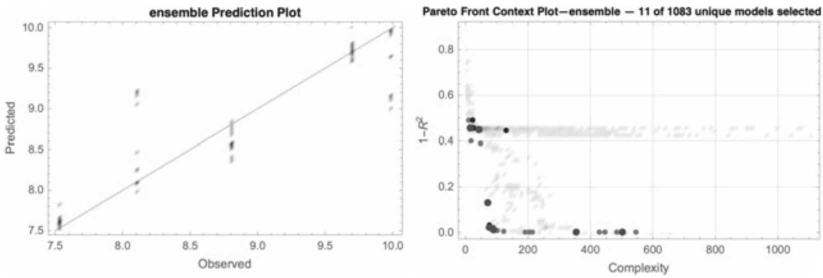


Fig. 4 Ensemble prediction and accuracy-complexity trade-off plots for sustainable exchange-traded funds (ETFs)

4.2 Sustainable ETFs

ETFs are alternative finance instruments based on traditional finance assets (stocks). As we have noted, sustainable ETFs are meant to provide exposure to ESG-oriented stocks. In Fig. 4, the prediction plot presents a categorical response prediction. In this model, only five numerical score values were supplied. For each of these observed scores, the predicted ranges, from 0.5 to more than 1, show significant differences which is a sign of uncertainty in the model. The Pareto plot shows improved error characteristics than the previous case of a categorical response modeling (lowest error around 0). However, the total population of models (outside the 11 included in the ensemble) has a more atypical shape, with a large concentration of those 1,083 models at error levels around 0.4. The model results show that just one driver variable is responsible for the ESG scores: *Carbon Intensity (Tons of CO2e/$M Sales).*

5 Discussion

5.1 Publicly Traded Private Equity Stocks

When comparing the ensemble models for publicly traded private equity stocks in Figs. 5 and 6, the ensemble2 residuals appear essentially free from modeling pathologies. Additionally, the residuals for the ensemble2 model make more evident the zones where we can expect more variability and less certainty about the results obtained.

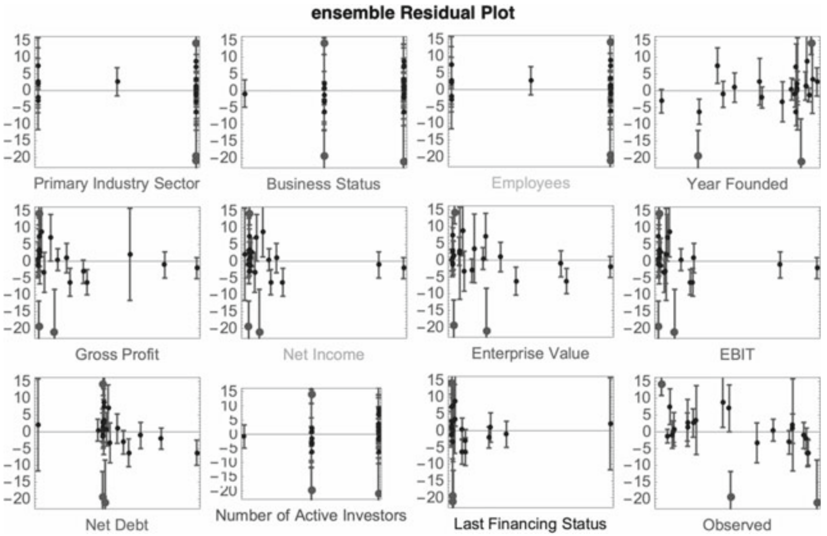


Fig. 5 Error residuals for the ensemble one model in the first round of modeling for publicly traded private equity stocks (*Note* The lighter areas are modeling outliers)

In the case of the variable *EBITDA* (earnings before interest, depreciation, and amortization), Fig. 7 suggests a high density of results and wider spreads in the response prediction for lower values of this variable. Additionally, we observe one outlier residual at the far right of the plot. This prevalence of extremes calls for caution when using this variable to assess ESG scores since there is not enough relevant information in the dataset to arrive at meaningful conclusions using that input. In this case, the main utility of using an ensemble model is not to predict the exact values of ESG scores but rather to help us identify what variables (and in which range of values) are needed to understand the score assigned to the financial asset.

Furthermore, we also discovered two driver variables common to both models by comparing the two ensemble models: *Employees* and *Year Founded*. This result suggests that both variables (which are signs of scale and maturity in a company) should be material in their effect on ESG scores in this particular industry.

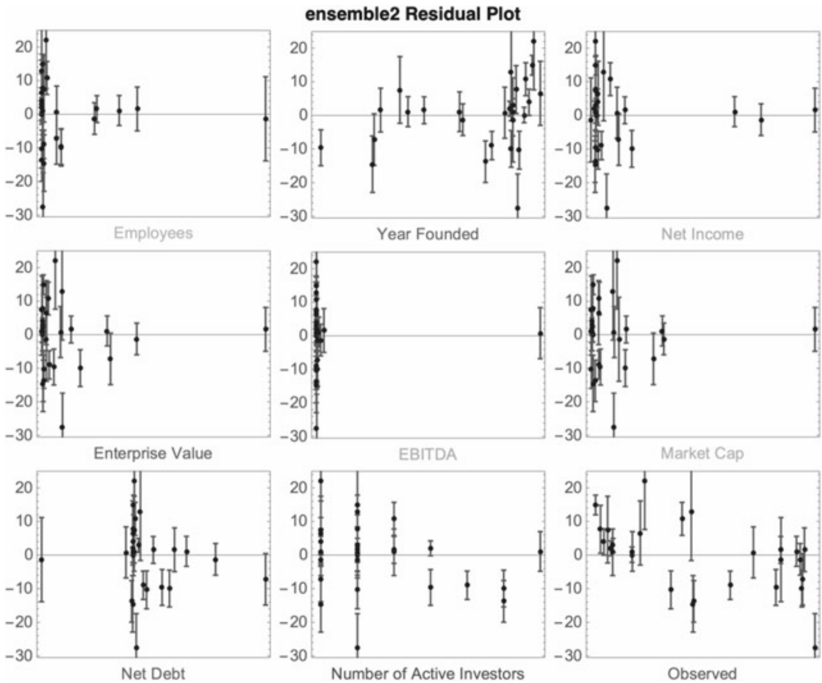


Fig. 6 Error residuals for the ensemble two model in the second round of modeling publicly traded private equity stocks

5.2 *Sustainable ETFs*

The fact that we obtained only one driving variable across all constituent models of the ensemble for sustainable ETFs is nontrivial. The ETF scores documentation clarifies that the current state of affairs in ESG rankings for ETFs is such that, in many cases, the only constituent of the overall ESG scores is a measure of environmental performance. Thus, it makes sense that *Carbon Intensity (Tons of CO2e / $M Sales)* appears as the prevalent variable in our models and functions as the indicator for the value of the target variable *ESG Score* in the ensemble model. In other words, it is to be expected by the reader that the ESG score will follow the carbon intensity of the underlying sustainable ETF, and the modeling process confirms that. This result is important for finance practitioners for several reasons. Firstly, there are ongoing initiatives to regulate the market for

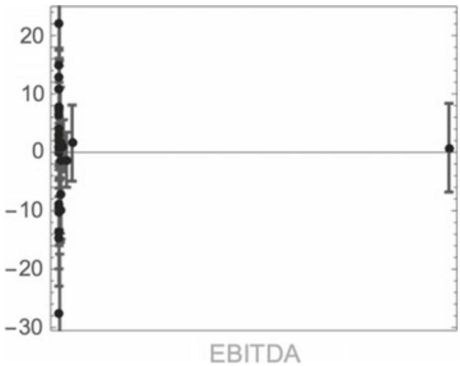


Fig. 7 Error residuals for the variable 'EBITDA' (earnings before interest, taxes, depreciation, and amortization) in the ensemble for the second round of modeling for publicly traded private equity stocks

ESG rankings, so providers may soon have to demonstrate the materiality of their assessments. Secondly, investors issue mandates and guidelines for their investment managers, and so those managers using ESG scores to support an investment thesis are also required to demonstrate how investment performance and ESG scores are related.

5.3 Ensemble Models Using Big Data

The growth of available data in the ESG investment space has the potential to impact the results obtained in this study because either incremental observations or new variables are likely to affect the choice of input variables for further modeling rounds in existing data. Fortunately, the genetic programming via symbolic regression technique is well suited to deal with large datasets with many variables. Big data can be used by researchers in this area to further clarify existing uncertainties by allowing the creation of models with data that might be available only within internal company data lakes. For example, a company may not fully disclose sensitive data such as debt figures to the public in real-time, but companies and their funders and partners may desire to assess the impact of capital structure changes on ESG scores and vice versa.

6 Conclusion

In conclusion, our chapter aimed to reduce uncertainty in the ESG investment space. We focused on the sustainability concerns of investors and new ones that reveal themselves through our modeling results. We first address uncertainty by providing data analysis of the relationship between investment-related variables and ESG scores. Investment managers currently do not have access to such analysis since providers only give them scores and sometimes the formulations they use to calculate scores. Thus, a more comprehensive review across groups of companies is not available. This chapter addresses this gap in the literature by suggesting a technique, symbolic regression using genetic programming, that makes those hidden relationships visible. The main conclusions for the ensemble models used for publicly traded private equity stocks and sustainable ETFs are detailed below.

6.1 Publicly Traded Private Equity Stocks

The reader cannot underestimate the utility of comparing ensemble modeling approaches for more mature and larger companies where more data is available. For instance, a large private equity investment company may have dozens of portfolio companies and hundreds of possible target investments, and thus the ensemble approach may help monitor and evaluate investment opportunities. When dealing with panel data as described above, it is helpful to identify redundancies that point to the true drivers and provide a greater awareness of uncertainty in those inputs. Finally, the ensemble fulfills its purpose by pointing out where to doubt the results, that is, where an investment manager is advised to perform additional due diligence.

6.2 Sustainable ETFs

Despite having a small number of observations (ETF symbols) in the dataset, generating the ensemble model was still informative. Explicit uncertainty ranges were obtained at a sufficiently high accuracy to be able to trust the measurement of uncertainty. However, other unexpected characteristics in the model, such as the shape of the Pareto front in Fig. 4, add to the other sources of uncertainty and deserve further study.

The goal of this chapter was to shine a light on a murky investment space, and we have provided some clarity with our modeling and analysis. We have highlighted where uncertainty exists and where we can find firmer ground on which to base our assumptions. Future research should look to refine our approach as access to more data becomes available.

Appendix

See Table 1.

Table 1 Data dictionary

Variables	*Definition*
ESG Score for stocks	A company ESG score is based on a company's reported information in the environmental, social, and corporate governance pillars (from Refinitv)
ESG combined score for stocks	The ESG score (above) with an ESG controversies overlay
MSCI ESG Quality Score	It measures the ability of underlying holdings to manage key medium to long-term risks and opportunities arising from environmental, social, and governance factors (for ETFs, from ETFdb)
Weighted Average Carbon Intensity	Tons of C02 emissions / $M sales. Measures the exposure of an ETF to carbon intensive companies
Number of Active Investors	The number of investors that actively hold a stake in the company after the most recent transaction of the company (from Pitchbook)
Employees	Count of current employees working at the company
Year Founded	The date the business was founded
First Financing Size	The amount of invested capital in the first financing event of the company
Last Financing Status	The stage of the deal at the time of the last financing event (e.g. complete, in progress, postponed, failed)
Web Growth Rate Percentile	The weekly percent change in the web signals of a company combines Similar web unique visitors and Majestic referring domains

(continued)

Table 1 (continued)

Variables	*Definition*
Primary Industry Sector	The high level principal business focus of a company in PitchBook's industry classification system
Business Status	The overall suspected cash flow of a company. E.g.: bankruptcy, generating revenue/not profitable, profitable
Gross Profit	The aggregate revenue less the cost of goods and services sold or operating expenses directly attributable to the revenue generation activity
Net Income	The net earnings of a corporation after deducting all costs of selling, depreciation, interest expense and taxes
Enterprise Value (EV)	The total value of a business. Calculated as (Market Capitalization + Total Debt + Minority Interest + Preferred Stock) – (Cash and Short-Term Investments)
EBIT	The earnings the company generated before paying interest and tax expenses
EBITDA	Earnings Before Interest, Taxes, Depreciation, and Amortization. Calculated as Revenue - Expenses (excluding interest, taxes, depreciation and amortization)
Market Cap	The market value of a company using most recently reported basic weighted average shares outstanding. Calculated as Stock Price * Basic Weighted Average Share
Net Debt	The overall debt situation of a company by netting the value of debts with cash and other similar liquid assets. Calculated as Short-term Borrowings + Long-term Debt – Cash & Cash Equivalents
Last Financing Status	Refers to the stage of the deal at the time of the last financing event (Ex. complete, in progress, postponed, failed)

References

Agudo, A. 2022. *Solving the ESG data conundrum* [Webinar]. Reuters. https://vimeo.com/681041950

Awasthi, S., & Muvija, M. (2017). *BAT warns of steeper declines in cigarette sales, shares slip*. https://www.reuters.com/article/us-brit-am-tobacco-outlook-idUSKCN1TD0NH

Brealey, R. A., Myers, S. C., & Allen, F. (2017). *Principles of corporate finance*. McGraw-Hill/Irwin.

Briand. R. (2022). What is environmental, social and governance? *MSCI*. https://www.msci.com/esg-101-what-is-esg

Cheong, M. (2022). *How ESG factors impact Credit Risk*. [Webinar]. S&P Global.

Eccles, R., & Serafeim, G. (2013). The performance frontier—Innovating for a sustainable strategy. *Harvard Business Review*. https://hbr.org/2013/05/the-performance-frontier-innovating-for-a-sustainable-strategy

Evolved Analytics LLC, DataModeler, Version 9.5, Rancho Santa Fe, CA USA. (2021). *The number of investors that actively hold a stake in the company after the company's most recent transaction*.

Fallender, S. (2020). *Quick chat on ESG disclosure: Through the lens of reporting*. Refinitiv [Webinar]. https://www.refinitiv.com/en/sustainable-finance/sustainability-perspectives/episode-41-through-the-lens-of-corporate-reporting.

Glover, J. (2021, October 15). *ESG rules and regulations will hamper business growth and wider societal change*. RSM UK. Insider.co.uk. https://www.insider.co.uk/news/esg-rules-regulations-hamper-business-25265333

IIGCC. (2019). *Global investor statement to governments on climate change*. https://www.iigcc.org/download/global-investor-statement/?wpdmdl=2600&refresh=625ed38bb70ad1650381707.

Kotin, A (host). Howard-Grenville, J. The Resonance Test. (2021). Episode number 65 [Audio podcast episode]. In *The Resonance Test*. Epam Continuum. https://soundcloud.com/futuremadereal/the-resonance-test-66-jennifer-howard-grenville

Kotanchek, M., Smits, G., & Vladislavleva, E. (2008). Trustable symbolic regression models—Using ensembles, interval arithmetic and Pareto Fronts to develop Robust and Trust-aware Models. In Rick L. Riolo, Terence Soule and Bill Worzel (Editors), Genetic Programming Theory and Practice V, chapter 12, Published 2008, ISBN: 978–0–387–76307–1.

Peterson, R., & Neuville, F. (2021). *Incorporating ESG Sentiment into Investment Strategies* [Webinar]. Refinitiv. https://solutions.refinitiv.com/ESGSentimentWebinar.

MSCI. (2020, December). *MSCI ESG ratings methodology*. https://www.msci.com/documents/1296102/4769829/MSCI+ESG+Ratings+Methodology+-+Exec+Summary+Dec+2020.pdf/15e36bed-bba2-1038-6fa0-2cf52a0c04d6?t=1608110671584

Rick, R., Ekaterina Vladislavleva, R., Ritchie, M. D., & Moore, J. H. (Editors). (2013, May). *Genetic programming theory and practice X. Series: Genetic and evolutionary computation*. ISBN 978–1–4614–6845–5, Hardcover. Springer.

Sand, C. (2021). *ESG is arriving faster than expected*. [Podcast]. Privcap Private Equity & Real Estate Podcast. https://open.spotify.com/episode/0VF7RZofs7TfmVDIvlvMou?si=8wqHrdHoTe2urMmCC3MDFQ&nd=1.

Stijven, S., Minnebo, W., & Vladislavleva, K. (2011). *Separating the wheat from the chaff: on feature selection and feature importance in regression random forests and symbolic regression*. In Steven Gustafson and Ekaterina Vladislavleva editors, 3rd symbolic regression and modeling workshop for GECCO 2011, pp. 623–630, Dublin, Ireland, ACM.

Securities and Exchange Board of India (SEBI). (2022). *Consultation Paper on Environmental, Social and Governance (ESG) Rating Providers for Security Markets.* https://www.sebi.gov.in/reports-and-statistics/reports/jan-2022/consultation-paper-on-environmental-social-and-governance-esg-rating-providers-for-securities-markets_55516.html

SECURITIES AND EXCHANGE COMMISSION. 2010. Volume 97 No 15. P. 2416. https://books.google.pt/books/about/SEC_Docket.html?id=LBk7AQAAMAAJ&redir_esc=y.

SECURITIES AND EXCHANGE COMMISSION. 2022. Release Nos. 33–11042; 34–94478; File No. S7–10–22. https://www.sec.gov/rules/proposed/2022/33-11042.pdf.

S&P Global. (2021). *The Sustainability Yearbook 2021*. https://www.upl-ltd.com/press_release/Fanzip28HTTQt9cyBTrxk31z9HQU9VtQ1reSgEXG.pdf.

Smits, G. F., Vladislavleva, E., & Kotanchek, M. (2011). *Scalable symbolic regression by continuous evolution with very small populations. In Genetic Programming Theory and Practice VIII*, Genetic and Evolutionary Computation Vol.8, Rick Riolo,Trent McConaghy, and Ekaterina Vladislavleva (Editors), Chapter 9, Published: 2011, Springer, ISBN: 978–1–4419–7746–5.

Snow, D. 2021. *ESG is arriving faster than expected*. [Podcast]. Privcap Private Equity & Real Estate Podcast. https://open.spotify.com/episode/0VF7RZofs7TfmVDIvlvMou?si=8wqHrdHoTe2urMmCC3MDFQ&nd=1.

S&P Global, How ESG factors impact Credit Risk.

The Global Compact. (2005). *Who cares wins: Connecting financial markets to a changing world*.

Thorne, J. (2021, Novomber 14). *At COP26, investors steal the stage from diplomats*. Pitchbook, Inc. https://pitchbook.com/news/articles/cop26-investors-venture-capital-clean-tech-climate

Venegas, P. (2021). Trustable risk scoring for non-bank gateways in blockchain and DLT financial networks. *Unifying Themes in Complex Systems X*.

Wheeler, M. (2019). *The future of Denim, Part #3: Waste not; water not—innovation*. Fashion Revolution. https://www.fashionrevolution.org/the-future-of-denim-part-3-waste-not-water-not-innovation/

Wolfram Research, Inc., Mathematica, Version 12.2.0, Champaign, IL (2021).

Big Data in Financial Services

Consumer Credit Assessments in the Age of Big Data

Lynnette Purda and Cecilia Ying

1 Introduction

People have been lending goods and currency for thousands of years, making it unsurprising that the format and process for lending has evolved over time. A key aspect of this process is the information upon which lending decisions are made. Blakstad and Allen (2018) suggest that banks have traditionally held a privileged information position since their interactions with consumers allow them to quickly access and verify information related to identity and transaction behavior. Based on this information, assessments of creditworthiness and lending decisions are made.

Today, non-bank sources of lending are increasing in both scale and innovation. Non-banks are already providing more home mortgages than banks (Dryer, 2020) and the expansion of non-bank lending across all other forms of consumer lending is increasing at a rapid pace. For

L. Purda (✉) · C. Ying
Queen's University, Kingston, ON, Canada
e-mail: lynnette.purda-heeler@queensu.ca; purdal@queensu.ca

C. Ying
e-mail: y.ying@queensu.ca

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_6

example, peer-to-peer lending by online platforms such as Lending Club and Prosper have attracted millions of borrowers since their inception in the mid-2000s and are providing increasingly unique inputs upon which investors can make their lending decisions.

In this chapter, we address what is driving the evolution in consumer lending with a particular focus on credit assessments. We explore two key aspects: (1) the information used in this assessment, and (2) the techniques that map this information to corresponding credit decisions. We illustrate the recent and considerable change in both aspects that has been driven in large part by big data and financial digitalization. Of specific relevance is the expanding scope of information that has been shown to be meaningful in establishing credit quality and the machine learning (ML) techniques that permit its analysis.

In order to better understand the evolution of consumer credit assessments, it is useful to provide an overview of traditional consumer credit scores and the challenges associated with these methods. These challenges have resulted in certain groups of consumers being underserved by financial intermediaries if they lack the required information. In many cases, the lack of information is a poor predictor of the individuals' ability or willingness to repay. As a result, alternative sources of data may serve as replacements for demonstrating credit quality. The expansion of lending from financial institutions to platforms, in combination with the availability of unique sources of information, has thus prompted the development of alternative credit assessment tools that may improve consumer access to funds and financial services. We review the wide range of possible data available including traditional Fair Isaac & Company (FICO) scores, hard financial information (such as income, indebtedness, car and home ownership status etc.), and the increasing use of soft information including self-reported data on borrower profiles, the purpose of the loan, and even social networks. We discuss the global reach of these changes and provide examples of how the unique needs of emerging markets have frequently prompted their development.

The inclusion of such varied and extensive data requires an increasingly sophisticated set of techniques to assess and analyze them. To a large extent, these techniques are applications of machine learning which Burns defines as "…a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so" (Burns, 2021). In the context of credit assessments, machine learning allows for the data to demonstrate how new and novel sources of information can be meaningful predictors of default.

Despite the potential for alternative data to boost financial inclusion and extend the reach of financial services around the globe, it is not without its challenges. Primary among these are concerns for privacy, the ethical use of consumer data, and the potential for systematic bias. We highlight these challenges and acknowledge that to be effective, the development of alternative consumer credit models must occur at the intersection of big data, machine learning, and ethics.

Our discussion of consumer credit assessments in the age of big data proceeds as follows. In the next section, Sect. 2, we provide a brief overview of traditional credit assessment data and techniques, followed by a discussion of the expanding scope of data that have been used as indicators of credit quality. This discussion, in Sect. 3, begins with the transition from traditional lenders to peer-to-peer platforms before moving to an examination of credit assessment from user-provided mobile phone data. Section 4 elaborates on the application of machine learning to consumer credit assessments and its integral role in facilitating the use of an ever-expanding world of data. Section 5 raises the issues of ethics and bias in both sourcing alternative data and the techniques used to analyze it. We conclude in Sect. 6 with a discussion of future research themes that may help inform best practices for consumer credit assessments.

2 Overview of Traditional Credit Assessment Data and Techniques

Historically, the creditworthiness of a borrower was determined based on the "4 'Cs' of credit" (Altman & Saunders, 1997), namely the borrower's character (reputation), capital (leverage), capacity (volatility of earnings), and collateral. These inputs, combined with the subjective judgment of a lending officer, determined who had access to additional funds. With the development of credit bureaus, such as TransUnion and Equifax, the process of consumer data collection moved away from the individual lender to the credit bureau, allowing for significant economies of scale. No longer did each individual lender need to collect and verify personal information; instead, this became the task of credit bureaus who, with the advent of computing technology in the 1970s and 1980s, became increasingly efficient (*FICO History|FICO*, 2022; *TransUnion Company History*, 2022).

Based on the aggregated data, credit bureaus developed their own unique consumer credit scores. While the precise composition of these scores is proprietary and varies from bureau to bureau, rough guidance on the factors that contribute to the scores is often provided. For example, Equifax reveals that FICO scores are determined from five main categories of consumer characteristics, namely a person's payment history, amounts owed, length of credit history, new credit accounts and types of credits used (Prod, 2022).

Once a credit bureau or potential lender has compiled the relevant information on borrower characteristics, they generally seek a way to aggregate it into a single score that summarizes overall credit quality. One of the first examples of this approach occurred in the context of corporate creditworthiness and the development of Altman's Z-score in 1968 (Altman, 1968, 2018). Altman suggested that five key financial ratios, namely working capital to assets, retained earnings to assets, earnings before interest and taxes to assets, book value of equity to liabilities, and sales to assets, could be mapped into a probability of financial distress. At issue was how to conduct this mapping and the weights assigned to each ratio. Altman established these through discriminant analysis, and a sample of firms that had gone bankrupt matched to those with similar characteristics that had not gone bankrupt. This analysis resulted in a set of coefficients for each of the five financial ratios as indicated below:

$$Z = 1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1X_5$$

where X_1 through X_5 correspond to the five financial ratios. At the time, Altman suggested that lower scores were more consistent with possible bankruptcy and that the model was capable of predicting default in the upcoming two years with an accuracy of 72%. The model was initially conceived in the context of public manufacturing firms but has since been extended to a variety of settings such as private firms and different industries.

Models of this format had obvious appeal. They were based on information that was widely available and easy to collect. Once the coefficients were established, it was trivial to translate the five variables into a single score. This marked the beginning of credit assessment models for both corporate and personal credit scores that derived the likelihood of loan repayments or loan defaults using a set of financial predictors. The resulting multivariate models fell into four main categories based on the underlying statistical techniques used: (1) linear regression (LR) models, (2) logistic regression/logit (LG) models, (3) probit models, and (4) discriminant analysis (DA) models (Altman & Saunders, 1997).

The LR models are used in credit scoring applications where the two-class problem (i.e., repaid versus non-repaid) is represented as a binary variable. The model classifies the borrowers into two groups based on a score calculated from a linear function constructed using continuous independent variables. In contrast, the logit and probit models produce a probability of non-repayment, rather than a simple classification. LG models assume the probability of default is logistically distributed instead of normally distributed as in the probit models. The DA models, such as Altman's Z-score described earlier, are conceptually similar to LR models, but use discrete instead of continuous independent variables for the classification.

Discriminant analysis, first introduced by Fisher (Fisher, 1936), emerged as the most popular among the four approaches for credit assessments (Abdou & Pointon, 2011) and has been a dominant methodology (Altman & Saunders, 1997), as it is simple to construct while producing a high prediction accuracy. Unfortunately, the technique also has some major disadvantages. First, it requires predictors to be independent and normally distributed. These characteristics are unlikely in the case of many consumer features which are frequently correlated. In addition, DA does not provide the significance of an individual predictor. For example, the relative strength of the predictive signals between two independent variables, say income and age, cannot be easily determined, making it difficult to understand the driving factors for predictions, specifically whether income or age is more important for repayment (Keasey & Watson, 1991). Another issue, although not unique to DA, is the use of prediction accuracy for measuring the model's performance, which ignores the unequal misclassification cost between Type I and Type II errors. In the context of credit assessments, a Type II error falsely predicts a future default when one does not occur. This is obviously less significant for the lender than a Type I error that misses to predict a default that does occur.

While ultimately the objective of a credit assessment should be predictive, such that it anticipates a consumer's future ability to pay, the categories used in measures such as FICO scores reflect a backward-looking assessment. This scoring becomes problematic for individuals without significant historical data, such as those new to obtaining credit, infrequent users of credit, individuals whose information may lie outside the reach of a domestic credit bureau, and the unbanked or under-banked with no significant financial record. Carroll and Rehmani (2017) estimate that approximately 75% of American adults have a credit score, leaving about 60 million individuals to be credit-invisible. The proportion with access to credit scores is not homogenous across the population, falling to 51% for African-Americans and 58% for Latinos (Carroll & Rehmani, 2017). Not having access to mainstream credit can make home ownership elusive, force individuals to use credit sources with excessively high-interest rates (such as payday loans), and effectively make it more expensive to borrow and thereby harder to get out of debt. Ironically, even those whose income preludes a regular need to borrow may be disadvantaged by traditional credit scores since the lack of historical data leads to "thin" or missing credit files. As a result, errors in credit assessments can cause certain segments of the population to be underserved since "in the end, the best credit scores go to good, regular users of credit, rather than customers who choose to take credit only when they can't avoid it" (King, 2014).

3 FinTech Lenders and Data Evolution

As lending moves beyond the reach of traditional financial institutions alone, it has changed the information available about prospective borrowers and consequently, the way in which their credit quality can be assessed. In this section, we examine how FinTechs have become a source of new credit assessment data and just how extensive the use of alternative data has become. We begin with a discussion of peer-to-peer (P2P) lending platforms, which form a useful bridge between traditional credit assessments and the incorporation of alternative data.

3.1 P2P Lending and Resulting Data Creation

Online lending platforms such as Zopa in the United Kingdom, and Prosper and LendingClub in the United States, have fundamentally changed the way that borrowers and lenders connect. Rather than requiring a financial intermediary such as a bank to pool surplus resources across lenders and then make allocation decisions to borrowers, lenders can now be much more involved in the lending decision. Although these peer-to-peer platforms may use credit bureau scores as one source of information, they frequently provide their own unique credit score for the loan as well. Giles Andrews, a co-founder of Zopa, suggests that the platform uses much of the same information as traditional lenders but will "buy more of it and use it more intelligently" (King, 2014), thereby lowering the default rate to less than what is typically seen for personal loans given by banks in the UK. By cutting out traditional financial intermediaries and moving the lending process online, Zopa can reduce overhead and become tremendously efficient, leading to lower interest expenses for borrowers.

The way that online lending platforms use credit bureau assessments in their process is not always clear and has changed over time. Studying the relation between FICO scores and LendingClub's proprietary ratings, Jagtiani and Lemieux (2019) document that the correlation between the two credit assessments has dropped from about 80% in 2007 to less than 35% in 2015. They argue that the decline in correlation has been driven by LendingClub's incorporation of additional metrics beyond inputs into FICO scores.

While it is unclear precisely what these alternative inputs are, some evidence exists that their incorporation improves credit assessments, particularly at the far ends of the credit scale. For example, some LendingClub loans that are assigned the best possible grade of "A" are from borrowers who maintain a low FICO score of less than 680. Despite this divergence between the proprietary grade and the FICO score, the default rate on these loans is in line with other A-rated credits with higher FICO scores, suggesting an ability to tease out those borrowers with a strong ability and willingness to pay despite a weak traditional credit score. The benefits for these borrowers are significant as their access to credit at lower rates than what would traditionally be provided is enhanced. At the other end of the spectrum, Jagtiani and Lemiueux (2019) document instances where LendingClub diverges from very high FICO scores, exceeding 750, to assign their lowest score of G. Default rates on these loans are significant, suggesting that the high FICO score may be flawed or at least stale and no longer relevant.

The unique credit scores that peer-to-peer lending platforms provide are one piece of information that a prospective lender can use when deciding whether to fund a loan. However, the platforms reveal additional information that is frequently "soft" in nature and provided by prospective borrowers. This information may describe details of the purpose of the loan or the characteristics of the borrowers themselves. For example, P2P platforms that allow the prospective borrower to provide a narrative describing the reason for the loan supply new textual and linguistic information that text mining and machine learning tools can use to identify patterns. As a result, the development of the FinTech itself, here the P2P platform, becomes a meaningful source of alternative data. Herzenstein et al. (2011) provide an example of the possible insights that can come from this data. Looking at the funding success rate, defined as how likely it is for the individual to receive the funds, they find that borrowers who provide a reason for their fund request are more likely to receive it. Similarly, Larrimore et al. (2011) suggest that the more specific the content of the request, the more likely the borrower is to obtain the loan. They find that extended narratives, concrete descriptions, and quantitative words all boost funding success. Moving from the persuasiveness of funding requests to associating language choice with credit assessments, Netzer et al. (2019) identify textual elements that augment the predictive ability for the loan's default beyond the use of financial and demographic variables alone. As a specific example, Gao, Lin, and Sias (2021) suggest that the readability of the funding explanation and the use of more positive words are associated with a lower probability of default.

3.2 The Expanding Scope of Alternative Data

As P2P platforms show, user-supplied data can complement or substitute for historical financial information that may be missing from a traditional credit assessment tool. A significant source of user data is also available from an individual's mobile phone. If a consumer agrees to provide access, it can supply a digital footprint for verifying identity and behavioral patterns, both of which can form essential inputs into credit assessments. Significant inroads have been made in emerging markets where mobile phone use may greatly exceed access to financial services. Below, we discuss innovations in credit assessments based on these alternative data sources.

LenddoEFL was first launched in the Philippines with the goal of providing credit to the emerging middle class. While many of these individuals are well educated and maintain steady employment, traditional credit providers often overlook them due to a lack of credit history. The credit scoring techniques developed by LenddoEFL take the perspective that "all data can be risk management data".[1] With their consent, the company taps into the information provided by borrowers' mobile phones and looks to establish a digital footprint. This information helps to establish and verify identity and can be used to develop an alternative credit score which the company suggests is based on indicators of personality, behavior, expertise, skills, and community.

Community plays a large role in the credit assessment model of LenddoEFL. The data provided by mapping relationships between various social media profiles such as Facebook, LinkedIn, and Twitter is used to aggregate information about friends, frequency of interactions, interests, and social networks. The company has also asked borrowers to identify a "Trusted Network" of individuals who essentially act as character references (Hardeman, 2012). To ensure some "skin in the game" for all participants, the credit scores of the trusted network will all suffer if the borrower fails to make payment on the loan. It is argued that this focus on community strengthens the incentives for repayment.[2]

How LenddoEFL disrupts the traditional process of credit assessments based on financial history is clear, as are the potential benefits for consumers in developing markets where FinTech adoption outpaces the global average. The Ernst & Young survey of FinTech adoption lists China, India, Russia, South Africa, and Colombia as all having over 75% of consumers reporting the use of FinTech, while the rate is only 50% in Canada and 46% in the USA.[3] Unsurprisingly, a host of alternative credit assessment tools have been developed to reach these potential consumers. Tala (https://tala.co/), with operations in Kenya, the Philippines, Mexico, and India, also provides borrowing decisions based on user smartphone data. Users apply directly from an app on their phone, and if approved, funds are immediately sent to their account. Branch (https://branch.co/) operates in a similar fashion and indicates that its credit scoring approach is based on details from a borrower's mobile handset, SMS logs, GPS data, contact lists, and repayment history. Using this data, it has provided over 21 million loans to consumers in Kenya, Tanzania, Nigeria, and India.

[1] https://lenddoefl.com/ourstory.

[2] A useful description of LenddoEFL's approach is provided by: https://digital.hbs.edu/platform-rctom/submission/its-all-about-who-you-know-lenddo-makes-credit-decisions-based-on-your-social-network/.

[3] https://www.ey.com/en_ca/ey-global-fintech-adoption-index.

What has become readily apparent is that the sources of alternative credit assessment data are virtually endless and the sheer volume of such data is huge. Clearly, the early statistical approaches such as discriminant analysis are incapable of keeping pace with such an exhaustive set of inputs. We turn next to exploring the innovations in techniques and tools that allow for the translation of this data into a meaningful consumer credit score.

4 Advancements in Methodologies and Technologies

Machine learning (ML) capabilities paired with cheaper and more accessible big data infrastructure, such as cloud computing, have enabled a shift toward more advanced modeling techniques for credit assessments. While we can process large volumes of data quickly, we can also extract relevant information from non-traditional sources and easily capture non-linear relationships between predictors to identify patterns.

Statistical and ML models can be categorized into two main groups: (1) parametric and (2) non-parametric (Clement, 2020). The advancement of computing capabilities promoted the use of non-parametric models, and the development of more advanced ML models have shown significant improvement in prediction performance. However, their prediction results are often difficult to interpret (Alaka et al., 2019). Nevertheless, a 2017 study that compared ML models to the traditional statistics techniques (DA and logistic regression) (Barboza et al., 2017) found that ML models outperformed traditional models by approximately 10% in producing more accurate default predictions. In the sections that follow, we discuss common techniques that can be used for both supervised and unsupervised ML and the resulting measures of accuracy. We highlight how these techniques can be applied in the context of credit assessments and the implications and limitations of doing so.

4.1 Common Classification Methodologies in ML

ML classification models can be categorized as supervised or unsupervised. For supervised algorithms, a target variable is provided, and the model learns from training data to predict unseen data. Popular supervised techniques include Decision Trees, Random Forest, Linear

Regression, Naïve Bayes, Gradient Boost, and AdaBoost. For unsupervised algorithms, no target variable is provided, and the model learns from the underlying structure of the data to group similar records into the same class. Popular unsupervised techniques for classification include various clustering techniques such as Hierarchical, DBSCAN, and K-means. In addition, there are two types of models that combine multiple algorithms (Louzada et al., 2016): (1) hybrid methods that combine different techniques to form the modeling and evaluation pipeline, for example, by backpropagating neural networks with traditional DA approach (Lee et al., 2002), and (2) ensemble methods (i.e., bagging, boosting, and stacking) which uses the outputs from one algorithm as inputs for the next, or uses a weighted output from multiple algorithms to form a more complex ensembled classifier. Hybrid models can sometimes be constructed as a type of stacking ensemble model.

In the context of consumer credit assessments and credit default prediction, both supervised and unsupervised techniques can be applied using data that captures key information related to the borrowers. For example, the borrower's age, address, income, occupation, list of assets (i.e., a home or car) and current liabilities (i.e., outstanding credit card balance, mortgages, or car loans) as well as the terms of the loan (i.e., amount, duration, repayment structure). This information is then engineered into predictors (referred to as features in ML). If the loan was repaid prior to maturity, it will be classified as non-defaulted and will be used as the target variable for supervised models. In an earlier review of 187 papers published between 1992 and 2015 (Louzada et al., 2016), hybrid and combined models were found to be the most popular, followed by Support Vector Machine (SVM) and Neural Network (NN). In a more recent review of 32 papers published between 2016 and 2020 (Clement, 2020), Logistic Regression was used as a baseline model in 20 of the 32 papers, compared with the implementations of SVM and NN as the second and third most popular algorithms tested.

Existing literature highlights that there is no one-size-fits-all credit methodology that is optimal and suitable for all credit assessment scenarios, and that "there is no ideal credit scoring modelling procedure that would guide the user in the choice of specific variables, cut-off score, validation method and sample size" (Abdou & Pointon, 2011). Different credit models have therefore been developed for specific applications, ranging from expert systems to simple statistical models to advanced ML models.

4.2 *Model Performance and Evaluation*

The Accuracy rate (ACC) has historically been an extremely popular performance evaluation metric for classification models. It measures the proportion of all correctly classified predictions as defaulted (True Positives, TP) and non-defaulted (True Negatives, TN) against all records. ACC is used in many review papers for comparing model performance (Abdou & Pointon, 2011; Clement, 2020; Louzada et al., 2016) and for evaluating new versus baseline models (Barboza et al., 2017; Papouskova & Hajek, 2019).

Despite its simplicity, ACC is a poor choice as a performance metric in the default prediction context. Default is a rare event, creating an imbalanced weight between default and non-default outcomes. Therefore, any model can achieve a higher ACC by prioritizing TN and predicting no defaults. Instead, the cost of the two types of misclassification errors should also be considered: (1) Type I error (False Positive, FP) for misclassified non-defaulted outcomes as defaulted and (2) Type II error (False Negative, FN) for misclassified defaulted outcomes as non-defaulted. ACC ignores the trade-off between these two types of misclassification errors and the asymmetric financial costs between them. Particularly, there is an actual financial cost for approving a loan that will later default, compared with an opportunity cost of lost business for rejecting a potentially good loan that will not default (Abdou & Pointon, 2011).

Given the imbalanced nature of default events, a comprehensive metric that accounts for the trade-off should be used in conjunction with ACC. Some common examples include Area Under the Curve (AUC), the Receiver Operating Characteristic curve (ROC Curve), sensitivity and specificity metrics, and F1-scores (Louzada et al., 2016). Sensitivity, also known as Recall or True Positive Rate, measures the ratio between the number of correctly classified positives (TP) cases to the total number of actual positive (TP + FN) cases (i.e., the proportion of predicted non-repayment that the model correctly classified). In contrast, Specificity, also known as True Negative Rate, measures the ratio between the number of correctly predicted negative (TN) cases to the total number of actual negative (TN + FP) cases. The F1-score has been viewed as the most comprehensive measure. It ties all components between Precision and Recall together to form a ratio between TP, FP, and FN. The ROC graph illustrates the relationship between the True Positive Rate (Sensitivity) and the False Positive Rate (1 – Specificity) across different thresholds. The AUC is then simply the area under the ROC. The AUC can be used to compare two algorithms with different ROCs based on the models' predictions, with a higher AUC indicating a better overall model with less FP or FN. Furthermore, Lohmann and Ohliger (2019) acknowledged the asymmetric financial costs between the two types of errors and proposed using the total cost of misclassification for evaluating the performance of the models.

5 Challenges, Biases, and Ethics

The combination of new data sources and more sophisticated techniques for their analysis has frequently been lauded as a way to improve consumer access to financial services while increasing efficiency and lowering borrowing costs. However, as with any innovation, these developments may have unintended consequences and can raise ethical concerns or issues of potential bias in lending decisions.

One such example of bias comes from user-provided details on P2P platforms. In addition to a user-supplied narrative or description of fund use, many online lending platforms allow borrowers to post additional information, including demographic details or even a photo. As in traditional lending involving a lending officer, there can be an element of subjectivity in the credit assessment and the potential for bias, either intentional or unconscious. Evidence suggests that these same issues occur in online lending. Pope and Sydnor (2011), Duarte et al. (2012), and Chen et al. (2017) all point to possible discrimination in online lending. In some cases, the source of bias is gender, while in others, it is driven by additional attributes revealed from the borrower's photo. These photos allow for the possibility that physical characteristics, such as ethnicity and appearance, influence lending decisions in an inappropriate and systematically biased way (Pope & Sydnor, 2011).

While borrower-provided narratives and photos give new information upon which lenders may (rightly or wrongly) base their decisions, the sources of information that can be gleaned about a prospective borrower from their mobile phone are far more extensive than this. What previously would have been irrelevant for a financial institution may now be used as indicators of character or behavioral patterns related to the probability of loan repayment. The extension to these sources of alternative data can raise questions of privacy and human rights and raise concerns as to who precisely is responsible for monitoring the use and security of personal information.

A report from Privacy International, a UK-based non-profit organization that seeks to study and advocate for the human right of privacy, alerts us to the potential dangers brought about by the expanding use of alternative data in credit assessments (Fintech, 2017). Most significantly, they point to three concerns. The first is that the alternative data used reveals details of individuals' lives that have never before been used in a financial context. Information such as text messages can become part of a credit profile and be used by financial services. A second related concern is that the decisions based on these alternative data sources change the nature of how financial decisions are being made by relying much more heavily on techniques such as machine learning. The consequences of decisions made using these techniques and how they may lead to bias is not always obvious. Third, the reliance on alternative forms of data such as people's location, social connections, and digital footprint may result in changes to individual behavior. While FinTech applications may undoubtedly promote positive financial behaviors, such as encouraging debt repayment or saving, they may also prompt individuals to alter their behavior in unintended ways.

The verification of identity through digital footprints and its consequences for precisely what is made visible also pose concerns around these alternative approaches. While a clear stumbling block to financial inclusion has been the absence of reliable identity indicators for large segments of the world's population, there are challenges associated with establishing identity through online profiles and social connections. Imagine the case of an individual who may hold political views that differ from a controlling regime. These individuals may go to great lengths to avoid establishing a significant online presence or, if they are active on social media, it is possible that their political views may be held against them when attempting to secure financial services. Similarly, social connections to religious or cultural groups, or association with the LGBTQ community may be used as potential basis for discrimination. Ganesh, Deutch, and Schulte (2016) suggest that marginalized individuals and their advocates may have different needs for privacy and security online.

While ethical concerns are raised when establishing the scope of what constitutes relevant data for assessing an individual's credit quality, the computational techniques used themselves raise issues of potential bias. ML models have undoubtedly provided some promising results and been shown to outperform traditional statistical models, but they have also inherited similar modeling challenges and introduced additional ML specific issues. Given the data imbalance where historically more loans are repaid than not, there are far fewer examples of non-repayment for training the models. Sample selection bias is another one of the biggest challenges for default prediction, since data collected usually only represents borrowers who have been approved and accepted the loan. This challenge means that the data is incomplete, with missing information related to borrowers who were rejected or who were approved but did not accept the loan. More specifically, even if the initial application information is available, the eventual outcome of default or non-default cannot be tracked.

Biased training data is another major data issue which leads to unfavorable outcomes for certain disadvantaged individuals or subgroups. In the ML fairness literature, three ways of correcting such bias have been proposed: (1) pre-processing, (2) in-processing, or (3) post-processing. Pre-processing adjustments are applied to the training data before it is fed into the model to achieve fair outcomes; in-processing adjustments are incorporated directly into the model; and post-processing adjustments are applied to the model's prediction instead. Although fairness in ML has become a rapidly growing research area, very few focus on fairness in the default prediction context. Kozodoi, Jacob, and Lessmann (2021) defined the statistical fairness criteria for credit scoring and examined the profit-fairness trade-off. They concluded that in-processors can reduce algorithmic discrimination to a reasonable level at a low cost. However, more work is needed to provide a standardized approach to measure and correct such biases specific to credit default predictions.

Hidden biases in ML models are yet another by-product of the increased model complexity and decreased model transparency, leaving users with inscrutable and unintuitive results (Selbst & Barocas, 2018). It is difficult for practitioners to trust and implement the more advanced ML models without a clear understanding of the key drivers behind the model's prediction. Luckily, many common ML classification models can be interpreted using standard packages such as SHAP (SHapley Additive exPlanations), and it should be incorporated as part of the standard credit default prediction model pipeline. However, this adds additional complexity to already implemented models, and model users will have to take precautions for any potentially unexpected findings like discriminatory model outputs.

6 Conclusions and Areas for Future Research

The previous discussion of alternative data, machine learning techniques used for its analysis, and both the benefits and challenges of consumer credit assessments using these tools allow us to draw the following conclusions. First, traditional credit scores based on historical financial data leave significant groups of consumers with less access to financial services than is warranted. In addition, the formulation of these scores may lead to perverse incentives. For instance, an individual with no particular need to borrow may in fact find themselves inclined to do so in order to show a consistent pattern of repayment. The additional financing costs therefore detract from the productive and efficient use of funds.

A second conclusion is that the use of alternative data holds significant promise for boosting financial inclusion for individuals if it can correctly identify behavioral patterns consistent with repayment. Access to additional funds at lower rates can support improvements in housing, education, and medical care for large segments of the population. However, there exist significant risks associated with the sharing of this information. Ownership for data security and privacy must be taken on by either the emerging FinTechs or the bodies that govern them. Opacity in precisely who is responsible for ensuring consumer safety in these regards is significant.

A final concluding observation is that the use of alternative data and ML techniques has not necessarily eliminated the subjective nature of credit assessments. At its origins, credit assessments relied on a judgment of an individual's character. Lending officers rely on their own heuristics, and the introduction of bias, whether knowingly applied or unconscious, has led to systematic discrepancies in how credit was provided that were frequently unrelated to an individual's true probability of repayment. While new data-driven ML techniques initially appear more objective in nature, we have seen that they pose an equally high risk of bias that is not yet well understood.

What becomes clear from these conclusions is that the scope for future research in the area of consumer credit assessments is significant. While there are undoubtedly many topics to be developed in the individual disciplines of finance, behavioral economics, data analytics, and computer sciences, what is urgently needed is a framework that integrates ethics, security, and privacy considerations related to the handling of these alternative credit assessment approaches. Apparent from the above discussion is a need for research that takes an interdisciplinary lens and forces considerations of bias and unintended consequences into the discussion. Only then can big data fulfill the promise of promoting equitable access to credit and financial inclusion.

References

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the literature. *Intelligent Systems in Accounting, Finance and Management, 18*(2–3), 59–88. https://doi.org/10.1002/isaf.325

Alaka, H., Oyedele, L., Owolabi, H., Akinade, O., Bilal, M., & Ajayi, S. (2019). A big data analytics approach for construction firms failure prediction models. *IEEE Transactions on Engineering Management, 66*(4), 689–698. https://doi.org/10.1109/TEM.2018.2856376

Altman, E. I. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance, 23*(4), 589–609. https://doi.org/10.2307/2978933

Altman, E. I. (2018). A fifty-year retrospective on credit risk models, the Altman Z-score family of models and their applications to financial markets and managerial strategies. *The Journal of Credit Risk*, *14*(4), 1–34. https://doi.org/10.21314/JCR.2018.243

Altman, E. I., & Saunders, A. (1997). Credit risk measurement: Developments over the last 20 years. *Journal of Banking & Finance, 21*(11), 1721–1742. https://doi.org/10.1016/S0378-4266(97)00036-8

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems with Applications, 83*, 405–417. https://doi.org/10.1016/j.eswa.2017.04.006

Blakstad, S., & Allen, R. (2018). *FinTech revolution*. Springer. https://econpapers.repec.org/bookchap/sprsprbok/978-3-319-76014-8.htm

Burns, E. (2021). *What is machine learning and why is it important?* SearchEnterpriseAI. https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML

Carroll, P., & Rehmani, S. (2017). Alternative data and the unbanked. *Oliver Wyman*, 17.

Chen, D., Li, X., & Lai, F. (2017). Gender discrimination in online peer-to-peer credit lending: Evidence from a lending platform in China. *Electronic Commerce Research, 17*(4), 553–583. https://doi.org/10.1007/s10660-016-9247-2

Clement, C. (2020). Macchine learning in Bankruptcy prediction—A review. *Journal of Public Administration, Finance and Law, 17*, 20.

Dryer, T. (2020). *Council post: The future of lending: Five trends to watch*. Forbes. https://www.forbes.com/sites/forbesfinancecouncil/2020/02/18/the-future-of-lending-five-trends-to-watch/

Duarte, J., Siegel, S., & Young, L. (2012). Trust and credit: The role of appearance in peer-to-peer lending. *Review of Financial Studies, 25*(8), 2455–2484. https://doi.org/10.1093/rfs/hhs071

FICO History|FICO. (2022). FICO history. https://www.fico.com/en/history

Fintech: Privacy and Identity In the New Data-Intensive Financial Sector. (2017, December 1). Privacy international. http://privacyinternational.org/report/998/fintech-privacy-and-identity-new-data-intensive-financial-sector

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Ganesh, M. I., Deutch, J., & Schulte, J. (2016). Privacy, anonymity, visibility: Dilemmas in tech use by marginalised communities. *Tactical Technology Collectiev*, 40.

Gao, Q., Lin, M., & Sias, R. (2021). Words matter: The role of readability, tone, and deception cues in online credit markets—JFQA: Journal of Financial and Quantitative Analysis. *Journal of Financial and Quantitative Analysis*, *Upcoming*. https://jfqa.org/2021/07/05/words-matter-the-role-of-readability-tone-and-deception-cues-in-online-credit-markets/

Hardeman, B. (2012, June 15). *Lenddo's social credit score: How who you know might affect your next Loan*. HuffPost. https://www.huffpost.com/entry/lenddos-social-credit-sco_b_1598026

Herzenstein, M., Sonenshein, S., & Dholakia, U. M. (2011). Tell me a good story and I may lend you money: The role of narratives in peer-to-peer lending decisions. *Journal of Marketing Research*, *48*(SPL), S138–S149. https://doi.org/10.1509/jmkr.48.SPL.S138

Jagtiani, J., & Lemieux, C. (2019). The roles of alternative data and machine learning in fintech lending: Evidence from the LendingClub consumer platform. *Financial Management, 48*(4), 1009–1029. https://doi.org/10.1111/fima.12295

Keasey, K., & Watson, R. (1991). Financial distress prediction models: A review of their usefulness1. *British Journal of Management, 2*(2), 89–102. https://doi.org/10.1111/j.1467-8551.1991.tb00019.x

King, B. (2014). A new take on credit and lending. In *Breaking Banks: The innovators, rogues, and strategists rebooting banking* (pp. 5–8). John Wiley & Sons.

Kozodoi, N., Jacob, J., & Lessmann, S. (2021). Fairness in credit scoring: Assessment, implementation and profit implications. ArXiv:2103.01907 *[Cs, q-Fin, Stat]*. http://arxiv.org/abs/2103.01907

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., & Gorski, S. (2011). Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research, 39*(1), 19–37. https://doi.org/10.1080/00909882.2010.536844

Lee, T.-S., Chiu, C.-C., Lu, C.-J., & Chen, I.-F. (2002). Credit scoring using the hybrid neural discriminant technique. *Expert Systems with Applications, 23*(3), 245–254. https://doi.org/10.1016/S0957-4174(02)00044-1

Lohmann, C., & Ohliger, T. (2019). The total cost of misclassification in credit scoring: A comparison of generalized linear models and generalized additive models. *Journal of Forecasting, 38*(5), 375–389. https://doi.org/10.1002/for.2545

Louzada, F., Ara, A., & Fernandes, G. B. (2016). Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science, 21*(2), 117–134. https://doi.org/10.1016/j.sorms.2016.10.001

Netzer, O., Lemaire, A., & Herzenstein, M. (2019). When words sweat: Identifying signals for loan default in the text of loan applications. *Journal of Marketing Research, 56*(6), 960–980. https://doi.org/10.1177/0022243719852959

Papouskova, M., & Hajek, P. (2019). Two-stage consumer credit risk modelling using heterogeneous ensemble learning. *Decision Support Systems, 118*, 33–45. https://doi.org/10.1016/j.dss.2019.01.002

Pope, D. G., & Sydnor, J. R. (2011). What's in a picture? Evidence of discrimination from prosper.com. *Journal of Human Resources*, *46*(1), 53–92. https://doi.org/10.3368/jhr.46.1.53

Prod, E. (2022). *What is a FICO® score, how is it calculated|Equifax—United States*. What is a FICO® score, how is it calculated|Equifax. https://www.equifax.com/personal/education/credit/score/what-is-a-fico-score/

Selbst, A. D., & Barocas, S. (2018). The intuitive appeal of explainable machines. *Fordham Law Review, 87*(3), 1085–1140.

TransUnion Company History. (2022). TransUnion company history. https://www.transunion.com/about-us/company-history

Robo-Advisors: A Big Data Challenge

Federico Severino and Sébastien Thierry

1 Introduction

The 2008 global financial crisis led to the belief that the growing development of digital technologies would gradually make traditional financial institutions disappear in favor of FinTech firms (Stulz, 2019). It is in this context of digital transformation that robo-advisors entered the market. In fact, these instruments are part of an already long-standing modernization trend in the financial industry. Phoon and Koh (2018) claim that the last half-century saw the increasing development of innovative technologies concerning security trading, transaction processing, and advisory services to investors. Digital advisors remarkably contribute to these areas. Specifically, these are defined by the Sovereign Wealth Fund Institute (2015) as "a type of financial advisor that provides web-based portfolio management with almost zero human intervention. These online advisors typically use algorithms and formulas to conduct portfolio management."

F. Severino (✉) · S. Thierry
Université Laval, Québec, QC, Canada
e-mail: federico.severino@fsa.ulaval.ca

S. Thierry
e-mail: sebastien.thierry.1@ulaval.ca

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_7

Appearing more frequently after the global financial crisis in 2008, the first robo-advisors essentially focused on passive portfolio strategies via exchange-traded funds (ETFs) and exploited algorithms to automatically rebalance positions. This approach, which reflected investors' prudent attitude at the time, led to the increased development of robo-advisors for passive asset management (Phoon & Koh, 2018).

As robo-advisors become increasingly important in the financial industry, several aspects of this novel technology should be investigated by researchers. On the one hand, many authors highlight the advantages offered by robo-advisors. For instance, Uhl and Rohner (2018) argue that many human managers do not sufficiently take into account the investment personality of individuals. It is essential that the proposed asset allocation matches the investor's attitude toward risk, especially for long-term investors. Robo-advisors could, thus, offer a higher degree of customization to investors.

On the other hand, several studies are rather critical of robo-advisors' simplicity and lack of sophistication. For example, Tertilt and Scholz (2018) show that the algorithms used to classify clients into different risk profiles often turn out to be too simplistic. They believe robo-advisors are currently in their initial stage of development, but that this technology has the potential to evolve. Furthermore, Fisch et al. (2019) observe that, given the over-simplified design of robo-advisors, more and more companies choose to combine the characteristics of robo-advisors with those of traditional human managers. Interestingly, the (partially) automated investment process of such "hybrid robots" induces a relevant reduction of fees while retaining the option for investors to communicate with a human portfolio manager.

Finally, digital advisors present other critical issues, as discussed in the literature. According to a report from the Financial Industry Regulatory Authority (FINRA, 2016), clients of robo-advisors are often left to determine whether the investment strategy proposed by the robot actually meets their needs. Fein (2015) also notes that robo-advisors take little account of the intrinsic characteristics of the client, which could be detrimental in the event of a major financial crisis.

In this chapter, we first illustrate the characteristics of robo-advisors, emphasizing their strengths and weaknesses. Then, we focus on how digitization and the use of big data can impressively improve the performance of automated asset managers. Indeed, big data analytics can boost the customization of the financial strategies offered by digital advisors. The latter could target investors with more specific needs and propose extremely tailored investment solutions thanks to the information extracted from social network data or the client's interaction with a chatbot. At its current stage, the application of such technologies to automated advisory is still embryonic in the financial industry, but its development potential is enormous, which we will discuss in this chapter.

2 Robo-Advisor Features, Benefits, and Drawbacks

Nicoletti (2017) lists robo-advisors among contemporary promising FinTech innovations that are reshaping the financial services industry (see Chapter 4 therein). In the following subsections, we illustrate the main characteristics of digital advisors by focusing on their advantages and disadvantages with respect to traditional human financial consultants. Table 1, at the end of the following section, summarizes the main points of our discussion.

2.1 Generalities and Recent Trends in the Financial Industry

Robo-advisors are becoming increasingly widespread globally, particularly in North America, Europe, and South-East Asia. A list of representative robo-advisory companies is provided in Table 1 of Xing et al. (2019), while Benson (2022) provides a snapshot of the most popular robo-advisors as of March 2022 in a Nerdwallet article.

Not all robo-advisors are equal, as some offer more advanced services than others. In a Deloitte (2016) report by Moulliet, Stolzenbach, Majonek, and Völker, robo-advisors are classified by the company into four broad categories. First- and second-generation bots are based solely on the use of an online questionnaire that serves as the basis for determining clients' risk levels. A traditional portfolio manager then fully defines and adjusts asset allocation. Third and fourth-generation robo-advisors are much more developed as they use quantitative methods and algorithms to build and rebalance portfolios. Unlike first- and second-generation machines, they perform fully automated portfolio management throughout the investment process. Figure 1 illustrates the details of these categories.

Figure 2 of Jung et al. (2018) summarizes the process of automated financial consulting by comparing it with the human alternative. The

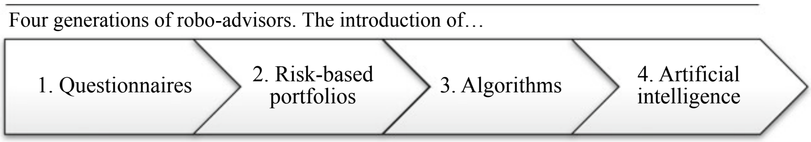


Fig. 1 The development of robo-advisors (*Note* More details in Deloitte (2016) report by Moulliet, Stolzenbach, Majonek, and Völker)

product configuration by robo-advisors replaces the traditional meeting with the client, profiling, and the definition of targets. The construction of a suitable portfolio is then automated by robo-advisors who match the client's profile with existing financial products (with a preference for passive investing through ETFs). Finally, the traditional portfolio rebalancing and the feedback to the client are also automatically implemented by digital advisors.

According to Phoon and Koh (2018), robo-advisors can be classified into three broad categories:

1. **Direct-to-consumer (D2C) model**: applies to online platforms that provide automatic portfolio management without human intervention,
2. **Business-to-business (B2B) model**: refers to platforms that help traditional financial advisors offer a digital wealth management solution, and
3. **Hybrid model**: includes personalized services for clients but also offers automated portfolio recommendations.

D2C models are probably the most iconic kind of robo-advisors, an example of these being the personal finance company SoFi. The service is provided by an application that a user can easily access via a smartphone. Due to product flexibility, the target clientele is highly diversified in terms of wealth, age, and financial needs.

Concerning the B2B model, Fisch et al. (2019) document a recent increased use of robo-advisors as a digital tool for human advisors. The automated advisory allows portfolio managers to guide their customers better and design the most appropriate investment strategy. This approach has been followed, for instance, by the Raymond James Financial group.

As for hybrid models, Fisch et al. (2019) acknowledge an increase in the hybridization of digital advisors in the United States as firms frequently combine the characteristics of digital advisors with those of traditional human managers. Hybrid robo-advisors require lower management fees because of their partially automated investment process. However, such hybrid bots also allow their clients to speak with a human advisor.

Hybrid robo-advisors increase competition between firms and create an appetite for heterogeneous types of clients. For example, the digital advisor Betterment offers three different levels of services. The first-level service is a standard robo-advisor, with no possible contact with a human advisor. The second-level service provides unlimited access to financial experts and licensed professionals. The third-level offer provides access to a dedicated advisor, in addition to automated consulting. The latter is attractive for mature, high-net-worth investors, whereas standard robo-advisor users are generally less wealthy and younger. This view is consistent with the digression of Weisser (2016) on Consumer Reports.

2.2 Robo-Advisor Benefits

One of the most remarkable advantages of digital advisors is the personalization (or customization) of the financial product they offer. As discussed in Chapter 1 of Sironi (2016), robo-advisors constitute the primary innovative tool in the panorama of personal finance, together with goal-based investing and gamification. The structure of digital advisors allows for the tailoring of financial investments and precisely targets potential clients (such as tech-savvy millennials). Such innovations, which are fostered by big data analytics and FinTech digitalization, are still under development and the related technological changes are progressing rapidly (see Sect. 3). In addition, in light of individualization, robo-advisors can be set to optimize after-tax returns and harvest tax loss, which are investor specific. This feature is also addressed by Jung et al. (2019).

Beyond customization, the advantages that robo-advisors offer over humans have been widely studied in the literature. In particular, Uhl and Rohner (2018) show that digital advisors have three main competitive advantages over traditional portfolio managers.

The first advantage concerns the optimal asset allocation proposed by robo-advisors according to the investors' goals and profiles. In order for investments to be successful, it is essential to design a robust and well-diversified asset allocation across a broad universe of assets, and it is crucial that such allocation perfectly match the investor's objectives and risk profile. Uhl and Rohner (2018) argue that, contrary to robo-advisors, many human portfolio managers do not sufficiently take into account the investment personality of individuals. For instance, a portfolio manager might decide to reduce risk during market downturns even though the client is not particularly risk-averse. This kind of decision can affect the probability of achieving the client's desired investment objectives.

The second comparative advantage of digital advisors is their cost. While traditional institutions mainly offer personalized investment strategies to their wealthier clients, robo-advisors can offer personalized solutions to less affluent clients too, thanks to their less expensive technology platform. As argued by Sharpe (1991) and empirically assessed by Fama and French (2010), active management is a zero-sum game on average before fees and even a negative-sum game after fees, and the performance of actively managed funds tends to be inconsistent over time. On the contrary, digital advisors rely on passive portfolio strategies, which are automated and, thus, much cheaper. Exact quantification of transaction cost saving in the digital advisory is difficult to obtain because of the discretion in the comparable alternative to be considered. For the same investment target, humans and robo-advisors would propose different asset allocations with different ex-post returns. However, since robo-advisors rely on passive investment, comparing the cost-effectiveness of passive versus active asset management can provide an indication of the amount saved. For instance, French (2008) estimates an outperformance of passive investing over the average of all active and passive investors, amounting to 67 basis points for the years 1980–2006. This performance corresponds to at least 10% of the expected real return on the U.S. stock market with due simplifications. Moreover, by focusing on robo-advisors, Uhl and Rohner (2018) approximate the total expense ratio (TER) of digital advisors to 0.5%, while the TER of human advisors is approximately 1.5%. This TER leads to an average gap in returns between a robo-advised strategy and a traditional one of roughly 4.4% per year, when a portfolio of 60% equity and 40% bonds is considered.

Finally, the third competitive advantage of robo-advisors is the absence of behavioral biases. The performance of individuals investing in stocks is often lowered by poor market timing and behavioral biases, such as overconfidence, mood, and emotion (see, e.g., Kahneman, 2011). Additionally, the home bias can also be relevant. A digital advisor does not suffer from such behavioral biases because it strictly and automatically rebalances its assets when the market rises or falls and uses the available data better than the average investor.

2.3 *Robo-Advisor Drawbacks*

Several studies in the literature are rather critical of the simplicity and the lack of sophistication of automated financial advisors. Among others, Tertilt and Scholz (2018) discuss the questionnaires used to define a client's attitude toward risk. In their sample of German, United States, and English robo-advisors, they show that clients are asked an average of 10 questions, but only 60% of the answers have a tangible impact on profiling. The algorithm that classifies clients into different risk profiles turns out to be too simplistic. In any case, digital advisors tend to propose conservative strategies with low-risk exposure. Moreover, they describe some variability in the composition of the proposed portfolios (for the same risk tolerance) depending on the robo-advisor's country. For example, U.S. digital advisors appear to create portfolios with a relatively high proportion of equities compared to German robo-advisors, especially for risk-averse clients. Boreiko and Massarotti (2020) draw similar conclusions. As a result, Tertilt and Scholz (2018) argue that, at present, robo-advisors lack the ability to categorize clients into different risk profiles and in the quality of their portfolio recommendations. Before replacing human asset managers, the algorithms of robo-advisors will have to be improved by using artificial intelligence and big data to create truly personalized portfolios, which we discuss in more detail in Sect. 3. The investment universe of digital advisors, which is still currently limited to ETFs, will also have to significantly expand.

The March 2016 Financial Industry Regulatory Authority (FINRA) report discusses many other critical points of digital advisors. First, the report reiterates that robo-advisors do not provide truly personalized investment advice. Users are often left alone to determine whether the investment strategy offered by the machine is truly relevant to their needs. Digital advisors are meant by their creators to be tools that allow customers to determine their risk tolerance and investment preferences. These tools allow the customer to subscribe to a recommended asset allocation for investors with similar preferences. Thus, it would be a mistake for investors to believe that digital advisors perfectly meet their actual financial needs. Second, the FINRA report mentions that robo-advisors do not eliminate conflicts of interest and could even worsen them. Pertaining to this concern, Fein (2015) shows that digital advisors use affiliate brokers, who may not offer the best price for customers. The profit resulting from this mechanism would be particularly high for robo-advisors who, on the contrary, insist on their low costs. The conflicts of interest found in the traditional financial industry can also be experienced by users of automated financial advice.

Jung et al. (2019) also call into question the fiduciary duty of digital advisors, i.e., their legal duty to act in the client's best interest. Based on Fein (2015), the authors note that robo-advisors give very little consideration to client idiosyncrasies or other external factors that could have serious consequences in the event of an extreme market decline. Fein (2015) concludes that robo-advisors are not meeting their fiduciary duty and that only well-trained portfolio managers can effectively manage portfolios.

Finally, Schwinn and Teo (2018) highlight the fact that digital advisors are subject to country-specific regulations in terms of investor protection, compliance requirements, and taxation rules. Such burdens have an impact on the geographic extension of automated advisory (especially in Asia) and make it challenging to attract international investors.

In conclusion, all of the abovementioned aspects demonstrate that human portfolio managers still have remarkable advantages over automated advisors to date. Table 1 summarizes the above-detailed advantages and disadvantages of robo-advisors.

3 Big Data and Artificial Intelligence in Robo-Advisory

In this section, before discussing the impact of big data analytics on automated asset management, we first discuss the possibilities of humanizing robo-advisor platforms. Then, we focus on the personalization of digital financial advice via big data. Lastly, we discuss the related algorithms implemented by robo-advisors.

Table 1 Main advantages and disadvantages of robo-advisors relative to human advisors

Advantages	*Disadvantages*
Asset management without human intervention	Investor are left alone to decide among different strategies
Customization of portfolio strategies	Strategies come from other (similar) investors
Investor's risk profile considered in portfolio optimization	Simplistic questionnaires for investor profiling
Specific investors targeted	Conservative strategies based on ETFs, are not always tailored
Wide range of investors	Low fiduciary duty on investors' idiosyncrasies
Cheap passive portfolio strategies	Conflicts of interest affecting strategy costs
Absence of behavioral and home bias	Country-specific strategies beyond investor's profile
Optimized after-tax returns and tax loss harvesting	Limitations from country-specific regulations

3.1 Humanization Inspired by Artificial Intelligence

An interesting issue for the success of digital advisory is the humanization of robo-advisors. The problem has its root in technological contexts, where artificial intelligence moderates the interaction between the machine and the user.

Several researchers question whether robo-advisors will be able to replace humans, or whether a hybridization of portfolio management will be more likely. For example, Hodge et al. (2021) ask whether giving a robo-advisor a human name will encourage investors to use it. They start from the observation that this practice is increasingly common in the field of new technologies. For instance, Apple named its virtual assistant *Siri*. Previous research into human–machine relationships has focused on simple interactions that any human could perform: asking Siri about the weather forecast is a relatively simple task. Conversely, financial consulting is more complex, and it is usually delegated to a professional. The existing literature on human–computer relationships may thus not apply in the case of digital advisors.

Interestingly, Hodge et al. (2021) find that investors are more likely to trust a robo-advisor with a name than one without a name when the robot performs a simple task. Nevertheless, the opposite is true when the robot is required to perform a difficult task. As a result, the benefits of robo-advisor humanization are still under discussion. Other ways to humanize robots may be considered, such as using avatars. In addition, the influence of other variables may be considered, including the investors' age and the gender attributed to the robot by its chosen name.

3.2 Big Data for Robo-Advisor Customization

As already mentioned in the previous paragraphs, a fundamental feature for the current and future success of automated asset management is customization, meaning, the personalization of the provided financial advice to the client's specific needs. In this regard, Fisch et al. (2019) document the increasing diversification of services offered by digital advisors, many of which offer new services that target particular cohorts of customers. For example, since women have a higher average life expectancy than men, the robo-advisor Ellevest offers different portfolios based on the client's gender. In addition, the robo-advisor True Link focuses on older investors as well as retirees. These kinds of customization are rather basic and do not exhaust the potential of automated consulting. The crucial point is the way in which digital advisors acquire and exploit investors' information. A fundamental contribution in this direction will be made through exploiting big data and artificial intelligence tools in the coming years.

For the purpose of this study, we rely on the definition of big data's 3Vs (i.e., volume, velocity, and variety) provided by Laney (2001). The amount of data, the volume, is often related to the number of concerned individuals, the observation frequency, and the number of observed features. The speed, or velocity, relates to data collection, processing, and updating. The variety refers to the heterogeneous nature of the data (quantitative, qualitative, longitudinal, and textual data, among others). Some examples are given by credit card purchases, financial transactions, online searches on search engines, activity on social networks, and location data, among others. The use of big data is pervasive in almost all industrial and service sectors. Interestingly, Xu (2021) mentions robo-advisors as an application of big data in the financial industry, in addition to risk control, customer value management, strategic marketing, credit risk assessment, and many others.

Faloon and Scherer (2017) shed some light on the quality of the investor information captured via the questionnaires and the ways in which such information is exploited in portfolio optimization, keeping in mind the client's investment goals. Some of the variables considered by robo-advisors for customer profiling are listed in Table 3 of Jung et al. (2019): income, investment horizon, liabilities, saving rate, age, gender, and degree of financial risk taken are some examples. Nonetheless, Faloon and Scherer (2017) conclude that robo-advisors are still far from providing customized investment solutions as they are still poorly individualized, a problem that financial institutions can mitigate by using the information extrapolated from big data. As discussed in the next section, multiple integrations of big data in digital advisory can be fruitful. Textual data extracted from the client's interaction with a chatbot can improve the effectiveness of questionnaires in profiling, even though there are current privacy issues concerning the storage of conversational data. Updated data about other investors' opinions and past choices can be clustered by the robo-advisor designer to recommend suitable investment solutions. Real-time information from social network activity can also be exploited by financial institutions that offer this service.

In a 2020 article for Forbes, Koksal reports the words of John Zhang, founder of the robo-advisor WealthGap, which unravel the forthcoming role of big data in the automated financial advisory: "[The] analysis of vast quantities of historical and financial data will uncover alpha opportunities that traditional analysis would otherwise overlook and give robo-advisors an edge over passive strategies and AI can process big data swiftly, allowing robo-advisors to adapt to changing market conditions and consumer behaviors much quicker in order to make better investment decisions. Time saved is key here."

3.3 Opening the Black Box

Xing et al. (2019) present two ways to increase robo-advisors customization via big data: dialogue and recommendation systems.

Dialogue systems include task-oriented systems and chatbots. The use of conversational data provided by chatbots is undoubtedly promising for automated asset management. However, some concerns, such as client privacy and data processing costs, cannot be overlooked. In addition, chatbots usually require a long interaction with the machine (a long dialogue history) to trigger the learning from textual data, which is not always feasible. The ultimate goal is to obtain a personalized dialogue system where the chatbot elicits the client's personality through regular interaction with them (see, e.g., Fung et al., 2016). This ability can be achieved by financial institutions by using some given personality models that allow the chatbots to grasp the user's attitude toward risk or uncertainty.

An illustration of the functioning of chatbots is found in Chen et al. (2017). Unlike task-oriented dialogue systems, chatbots can hold a conversation with the user on diversified topics. Chatbots based on generative models can produce appropriate replies that have never appeared before. Among them, sequence-to-sequence models associate a response sequence (in words) to a given input sequence (always in words) through an encoder-decoder structure with a hidden context. The encoder uses a recurrent neural network to read the input sequence and map it into a context vector, while the decoder maps the context vector to the response sequence, also through a recurrent neural network. An optimization problem is solved to determine the most likely appropriate answer. Day, Lin et al. (2018) have recently attempted to integrate conversational data into the asset allocation model of digital advisors. The authors introduce a knowledge-based and generative-based dialogue system into the robo-advisor architecture to moderate the interaction via a sequence-to-sequence model.

Robo-advisors can design multiple investment strategies and assign whichever is best suited to the client based on their investment targets and risk tolerance; this is referred to as a recommendation system. To determine suitable products, the system exploits the customer's personal information, previous investment choices, and the behavior of other investors with similar profiles and opinions. In these steps, big data is essential, and numerous machine learning methods serve as effective tools. Unsupervised learning methods (e.g., k-means or agglomerative clustering) allow one to find groups of observations with homogeneous features, while principal component analysis helps with dimension reduction. Supervised learning methods (e.g., random forests, artificial neural networks, and many others) are useful for classification and regression problems (see, for instance, James et al., 2013). These tools allow for customer profiling and segmentation. Moreover, Xue et al. (2018) designed a robo-advisor that exploited investors' social relations and provided group recommendations based on financial social networks.

As Schwinn and Teo (2018) described, a successful example of the use of big data and artificial intelligence in automated advisory is that of the Asian mobile trading and investing service 8 Securities (now known as SoFi Hong Kong). The platform provided by this company has access to news and social media feeds. The related mobile app is effectively a social trading portal where trading information can be shared among peers. In 2016, the company introduced the robo-advisor Chloe. As reported on Finews.asia on August 1, 2016: "Powered by artificial intelligence and machine learning technologies developed in-house, Chloe will learn day by day as its user base and database grow to optimize goal-setting and portfolio matching for customers with different financial needs." By using machine learning techniques, Chloe becomes increasingly able to predict a client's investment targets and desired savings. Moreover, the portfolio performance can be checked directly from the investor's smartphone. In doing so, SoFi can target millennials with fragmented wealth who, as opposed to high-net-worth investors, do not usually turn to financial advisors.

Finally, in addition to improving customization, big data can be directly incorporated into the optimization algorithms used by digital advisors. Beketov et al. (2018) list the most popular portfolio optimization methods employed by robo-advisors. Modern Portfolio Theory turns out to be the most widespread framework for portfolio optimization still today. However, robo-advisors also implement numerous improvements of Markowitz (1952) to reduce extreme portfolio weight sensitivity and high data dependence, and to include higher moments, tail risks, and various optimization constraints. A first attempt to exploit big data in this direction is given by Day, Cheng et al. (2018). The robo-advisor proposed by them adopts the Black and Litterman (1992) approach, which includes the investor's subjective view on expected returns in the portfolio optimization problem. Such a subjective view is extracted from big data analysis from diversified sources, including the investor's characteristics, updated asset prices, financial forecasts, and so on. A learning module from these sources is integrated into the overall architecture of the digital advisor.

4 Conclusion

In conclusion, this chapter briefly discussed the advantages and disadvantages of robo-advisors, a disruptive technology for the financial industry. Although digital advisors aim to provide cheap personalized investment solutions to a wide range of clients, existing literature highlights their limitations in reaching this goal. However, this technology can highly profit from the analysis and treatment of big data.

More research must be done by the research community to properly capture investors' preferences and behavioral peculiarities for the future development of automated advisory through big data analytics. As outlined by Jung et al. (2018), future robo-advisors are expected to provide truly customized investment solutions, in a fully automated way, by employing a widespread user-friendly technology. Moreover, Xing et al. (2019) acknowledge that the influence of investors' personalities on their perception of risk is largely overlooked by the dialogue systems that aim at modeling them.

Other improvements are also expected in the functioning of dialogue systems. For instance, Chen et al. (2017) warn that conversational data on specific domains can be scarce. Dialogue agents should be able to learn autonomously, and elaborate concepts extracted from the web and become smart tools that rely less on repetitive training. Another delicate aspect of managing is privacy protection because many clients interact with the same dialogue agent, so confidentiality must be granted.

Quantitatively, we expect the use of advanced machine learning and optimization methods by digital advisors to become widespread in the near future (Beketov et al., 2018). This approach meets the demands of investors, which are very demanding in terms of technology and sophistication. The marketing potential of such a quantitative direction is enormous.

Finally, future developments in the exploitation of big data for automated advisory must address the issue of unfairness and bias in machine learning algorithms. As illustrated by Mehrabi et al. (2021), behavioral biases can be present in the human-created sample data used for training the algorithms. The algorithms can amplify such distortions and produce biased outcomes. When financial institutions integrate financial and social network data into digital advisors, such biases need to be considered to avoid inappropriate investment decisions. Mehrabi et al. (2021) list possible ways to mitigate this problem in many contexts, but the search for effective solutions is not over.

Financial and social network big data has the potential to improve the personalization of digital advisory in terms of customers' profiling and investment solution recommendations. Conversational data generated from the clients' interaction with a chatbot can help elicit their personality and financial needs. In addition, sophisticated machine learning methods integrated into the portfolio optimization algorithms used by digital advisors are promising tools for the future of financial consulting.

References

Beketov, M., Lehmann, K., & Wittke, M. (2018). Robo advisors: Quantitative methods inside the robots. *Journal of Asset Management, 19*(6), 363–370. https://doi.org/10.1057/s41260-018-0092-9

Benson, A. (2022, February 28). 11 best robo-advisors of March 2022. *Nerdwallet*. https://www.nerdwallet.com/best/investing/robo-advisors

Black, F., & Litterman, R. (1992). Global portfolio optimization. *Financial Analysts Journal, 48*(5), 28–43. https://doi.org/10.2469/faj.v48.n5.28

Boreiko, D., & Massarotti, F. (2020). How risk profiles of investors affect robo-advised portfolios. *Frontiers in Artificial Intelligence, 3*, article 60. https://doi.org/10.3389/frai.2020.00060

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter, 19*(2), 25–35. https://doi.org/10.1145/3166054.3166058

Day, M.-Y., Cheng, T.-K., & Li, J.-G. (2018). AI robo-advisor with big data analytics for financial services. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 1027–1031). IEEE. https://doi.org/10.1109/ASONAM.2018.8508854

Day, M.-Y., Lin, J.-T., & Chen, Y.-C. (2018). Artificial intelligence for conversational robo-advisor. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* (pp. 1057–1064). IEEE. https://doi.org/10.1109/ASONAM.2018.8508269

Faloon, M., & Scherer, B. (2017). Individualization of robo-advice. *The Journal of Wealth Management, 20*(1), 30–36. https://doi.org/10.3905/jwm.2017.20.1.030

Fama, E. F., & French, K. R. (2010). Luck versus skill in the cross-section of mutual fund returns. *The Journal of Finance, 65*(5), 1915–1947. https://doi.org/10.1111/j.1540-6261.2010.01598.x

Fein, M. (2015). Robo-advisors: A closer look. Available at SSRN. https://doi.org/10.2139/ssrn.2658701

Finews.Asia. (2016, August 1). *Hong Kong robo-advisor set to democratise global investing*. https://www.finews.asia/finance/22726-robo-advisor-fintech-asia-8-securities

FINRA. (2016). *Report on digital investment advice*. https://www.finra.org/sites/default/files/digital-investment-advice-report.pdf

Fisch, J., Laboure, M., & Turner, J. (2019). The emergence of the robo-advisor. In J. Agnew & O. S. Mitchell (Eds.), *The disruptive impact of FinTech on retirement systems* (pp. 13–37). Oxford University Press. https://doi.org/10.1093/oso/9780198845553.001.0001

French, K. R. (2008). Presidential address: The cost of active investing. *The Journal of Finance, 63*(4), 1537–1573. https://doi.org/10.1111/j.1540-6261.2008.01368.x

Fung, P., Dey, A., Siddique, F., Lin, R., Yang, Y., Bertero, D., Wan, Y., Chan, R., & Wu, C.-S. (2016). Zara: A virtual interactive dialogue system incorporating emotion, sentiment and personality recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations* (pp. 278–281). https://aclanthology.org/C16-2058

Hodge, F., Mendoza, K., & Sinha, R. (2021). The effect of humanizing robo-advisors on investor judgments. *Contemporary Accounting Research, 38*(1), 770–792. https://doi.org/10.1111/1911-3846.12641

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning with applications in R. *Springer*. https://doi.org/10.1007/978-1-4614-7138-7

Jung, D., Dorner, V., Glaser, F., & Morana, S. (2018). Robo-advisory: Digitalization and automation of financial advisory. *Business & Information Systems Engineering, 60*(1), 81–86. https://doi.org/10.1007/s12599-018-0521-9

Jung, D., Glaser, F., & Köpplin, W. (2019). Robo-advisory: Opportunities and risks for the future of financial advisory. In V. Nissen (Eds.), *Advances in consulting research* (pp. 405–427). Springer. https://doi.org/10.1007/978-3-319-95999-3_20

Kahneman, D. (2011). *Thinking, fast and slow*. Ferrar, Straus and Giroux.

Koksal, I. (2020, April 18). How AI is expanding the applications of robo advisory. *Forbes*. https://www.forbes.com/sites/ilkerkoksal/2020/04/18/how-ai-is-expanding-the-applications-of-robo-advisory/?sh=77dcfdea55c3

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META Group Research Note, 6*(70)

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance, 7*(1), 77–91. http://www.jstor.org/stable/2975974

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Moulliet, D., Stolzenbach, J., Majonek, A., & Völker, T. (2016). The expansion of robo-advisory in wealth management. *Deloitte*. https://www2.deloitte.com/content/dam/Deloitte/de/Documents/financial-services/Deloitte-Robo-safe.pdf

Nicoletti, B. (2017). *The future of FinTech. Integrating finance and technology in financial services*. Palgrave Macmillan, Springer. https://doi.org/10.1007/978-3-319-51415-4

Phoon, K., & Koh, F. (2018). Robo-advisors and wealth management. *The Journal of Alternative Investments, 20*(3), 79–94. https://doi.org/10.3905/jai.2018.20.3.079

Schwinn, R., & Teo, E. (2018). Inclusion or exclusion? Trends in robo-advisory for financial investment services. In *Handbook of blockchain, digital finance, and inclusion, Volume 2* (pp. 481–492). Elsevier. https://doi.org/10.1016/B978-0-12-812282-2.00021-8

Sharpe, W. (1991). The arithmetic of active management. *Financial Analysts Journal, 47*(1), 7–9. https://doi.org/10.2469/faj.v47.n1.7

Sironi, P. (2016). *FinTech innovation: From robo-advisors to goal based investing and gamification*. Wiley.

Sovereign Wealth Fund Institute. (2015, August 27). *Everyone wants a robo-advisor, right?* https://www.swfinstitute.org/news/34448/everyone-wants-a-robo-advisor-right-8461344

Stulz, R. (2019). Fintech, bigtech, and the future of banks. *Journal of Applied Corporate Finance, 31*(4), 86–97. https://doi.org/10.1111/jacf.12492

Tertilt, M., & Scholz, P. (2018). To advise, or not to advise—How robo-advisors evaluate the risk preferences of private investors. *The Journal of Wealth Management, 21*(2), 70–84. https://doi.org/10.3905/jwm.2018.21.2.070

Uhl, M., & Rohner, P. (2018). Robo-advisors versus traditional investment advisors: An unequal game. *The Journal of Wealth Management, 21*(1), 44–50. https://doi.org/10.3905/jwm.2018.21.1.044

Weisser, C. (2016, July 28). The rise of the robo-adviser. *Consumer Reports*. https://www.consumerreports.org/personal-investing/rise-of-the-robo-adviser/

Xing, F., Cambria, E., & Welsch, R. (2019). Intelligent asset management. *Springer Nature Switzerland*. https://doi.org/10.1007/978-3-030-30263-4

Xu, J. (2021). A comprehensive survey on applications of big data in the financial industry. In R. Thurasamy, I. A. Majid, N. A. B. A. Rahman & A. H. Zulkafli (Eds.), *Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021)* (pp. 311–315). Atlantis Press. https://doi.org/10.2991/assehr.k.211209.053

Xue, J., Zhu, E., Liu, Q., & Yin, J. (2018). Group recommendation based on financial social network for robo-advisor. *IEEE Access, 6*, 54527–54535. https://doi.org/10.1109/ACCESS.2018.2871131

Bitcoin: Future or Fad?

Daniel Tut

1 Introduction

Despite the worldwide attention that Bitcoin (and cryptocurrencies in general) has garnered, there are still significant uncertainties regarding the nature and the role of Bitcoin in an individual's portfolio. This chapter, using monthly data obtained from Coinmarketcap.com, Federal Reserve Bank St. Louis (FRED), and Yahoo finance for the period 2009–2022, examines whether Bitcoin is a commodity, a currency, an investment, a collectible, a store of value, or a cash proxy as well as its potential role in a person's portfolio. It also discusses how the Blockchain technology underpinning Bitcoin's protocol can be used to address some of the challenges in big data.

Bitcoin is a digital currency that uses a blockchain protocol in which digital signatures are cryptographically validated via timestamp, independent of a financial intermediary. A hash function that links transactions of any arbitrary size to a fixed value size is used to chain the timestamped

D. Tut (✉)
Ted Rogers School of Management, Toronto Metropolitan University, Toronto, ON, Canada
e-mail: dtut@ryerson.ca

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_8

blocks together. A hashcash proof-of-work (PoW) algorithm validates all transactions and blocks, creating a verifiable distributed timestamp digital ledger.[1,2] The blockchain technology underlying Bitcoin's protocol has a potentially wide range of applications and has been used to create new digital assets. Blockchain technology also has the potential to address some of the challenges underlying big data. The challenges in managing, collating, and storing big data arise due to size, frequency, complexity, data breaches, and computational needs. Blockchain technology can be used to create immutable protocols which minimize malicious attacks and fraudulent activities (Foley et al., 2019). Furthermore, due to the decentralized nature of blockchain protocols, owners can have control rights over the use of their private data independent of any third party. We provide evidence of how blockchain technology can be used to address some of the big data challenges in the healthcare sector, in the global supply chains, and in the control over access and use of private data.

Bitcoin's initial goal was to be a peer-to-peer electronic cash, a potential replacement and alternative to fiat currency. However, this chapter demonstrates that Bitcoin fails to function as an effective unit of account since transacting parties will always revert to using cash over Bitcoin in a universe in which cash exists. Bitcoin has lower scalability and higher transaction costs than alternative payment processes such as credit cards. Furthermore, because of its inherent price volatility, Bitcoin does not fair better than gold as a store of value, inflation hedge, and currency hedge. Bitcoin has a high Sharpe ratio, and its speculative nature can provide some alpha opportunities for institutional investors and hedge funds. However, given its speculative nature, it is not clear yet what role Bitcoin can serve in a regular person's portfolio.

The rest of this chapter is organized as follows. Section 2 examines whether Bitcoin is the future of payment systems; Sect. 3 discusses whether blockchain technology can address some of the challenges in big data and the potential impact of regulations on the cryptocurrency space; Sect. 4 concludes the chapter.

[1] See Nakamoto (2008), for detailed discussions on Bitcoin's protocol.

[2] Node operators or "miners" are incentivized and rewarded in Bitcoin for every successfully validated block in the chain. This ensures continuity of the Blockchain.

2 Is Bitcoin the Future of Payment Systems?

2.1 Bitcoin as a Cash Proxy

Is Bitcoin "money?" Bitcoin was originally intended to be a "purely peer-to-peer (p2p) version of electronic cash…[to] allow online payments to be sent directly from one party to another without going through a financial institution…. [and] a solution to the double-spending problem using a peer-to-peer network" (Nakamoto, 2008). In order to understand whether Bitcoin can actually function as "money," we need to examine whether it satisfies and meets the criteria for an item or for an object to be considered as such. There are four such attributes of money[3] (1) Medium of exchange, (2) Method of payment, (3) Unit of account, and (4) A store of value.

Bitcoin meets the first two criteria as the Bitcoin protocol allows for transactions using BTC, which is the smallest tradable unit of Bitcoin, to be transferred from one account to another. New transactions are communicated to nodes, and each node collects all transactions into blocks. Once a node finds a proof-of-work (PoW), the block is then communicated to all nodes. The block is only accepted if transactions are verified as having not been already spent. Once transactions are verified and validated, the nodes start working on creating and adding a new block to the chain. Transactions are only considered valid after they are verified through a community consensus; that is, by the majority of the network nodes (Akcora et al., 2018; Song & Aste, 2020). The network rejects any transaction whose referenced output does not exist or has already been spent; such a transaction is not included in the blockchain. In creating the block, a transaction is only added to the wallet if the sum of the block creation fee and transaction fees are greater than the Coinbase value (Easley et al., 2019). Matching transactions are then deleted from the pool before the block is relayed to peers and added as part of the main branch in the chain via a Merkle tree.

Effectively, Bitcoin eliminates double-spending via the use of a digital signature algorithm and a proof-of-work via a hash function, which provides some security and allows users to engage in exchanges (Lipton & Treccani, 2021). Thus, Bitcoin can serve as a medium of exchange and payment method (Yermack, 2015; Almudhaf, 2018; McLeay et al., 2014).

[3] Smithin, 2002; Davies, 2010; Goetzmann & Goetzmann, 2017; Keynes, 2018

Indeed, Fig. 1 shows that the price, volume, and market capitalization reflect the demand and interest in Bitcoin. This figure indicates that Bitcoin satisfies the first two of four attributes.

However, at best, Bitcoin partially meets the third criteria: "Unit of account." For Bitcoin to be a stable and effective unit of account, transacting parties should be able to price goods in Bitcoins. The daily fluctuations in Bitcoin prices, as shown in Fig. 2, suggest that it might neither be in the best interest of the buyer nor the seller if goods are priced in Bitcoin. For example, consider a one-time transaction between a buyer and a seller. If the value of Bitcoin is precipitously falling, the buyer might be willing to exchange their Bitcoin holdings for a basket of goods, while the seller will be unwilling to accept Bitcoin as a form of payment. Moreover, if the value of Bitcoin is on the rise, then buyers will find it difficult to depart with their Bitcoin holdings, yet this is precisely the time during which sellers are more than willing to accept and price goods in Bitcoin.

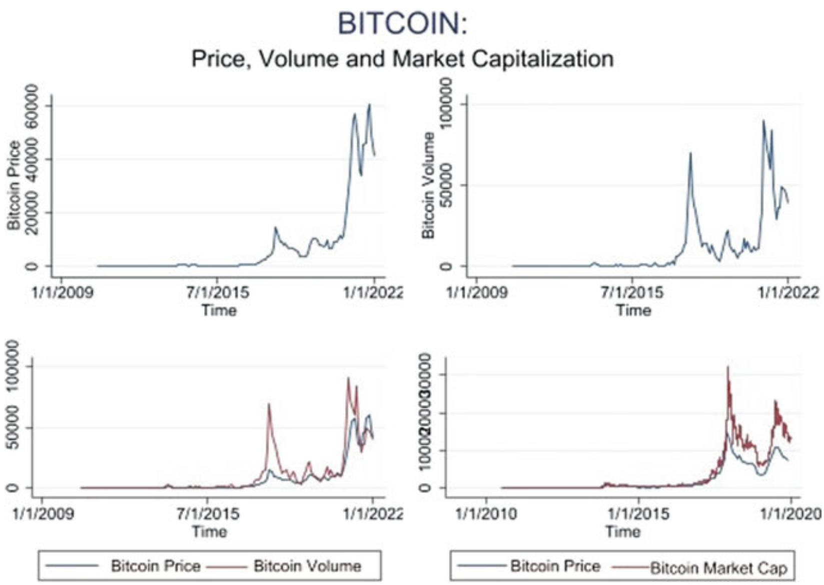


Fig. 1 Price, volume, and market capitalization of Bitcoin

Fig. 2 Price volatility of Bitcoin

This simple example illustrates that transactions are likely to be incomplete when goods are priced in Bitcoin. The inefficiency in Bitcoin transactions becomes even more apparent when considering issues in the labor markets (wages) and the financial markets (earning reports). Because of this inefficiency, it is difficult to make forward-looking valuations and to engage in future contracts when goods are priced in Bitcoin. Relative to alternative forms of payment, such as cash and credit cards, Bitcoin has higher transaction costs, as the mining of tokens is costly, and users have to utilize exchanges to receive tokens before engaging in any transactions (Stoll et al., 2019; Thum, 2018). The exchanges, such as Coinbase and Binance. US, serve as trusted third parties in the network. Additionally, each BTC block is limited to a 1 megabyte (MB) size and cannot handle more than eight transactions per second.[4] This restriction limits the scalability and broader adoption of Bitcoin as a form of payment on a global

[4] Credit cards such as Visa can process about 20,000 transactions per second with significant less energy consumption per transaction.

scale.[5] As a result, in a universe where cash exists, rational employers and employees would likely revert to using cash over Bitcoin.

2.1.1 Stablecoins

One potential solution to the volatility of Bitcoin as a currency is the advent of stablecoins. Stablecoins are digital currencies whose value is pegged to a fiat currency (U.S. dollars) or a basket of currencies. The aim is to use blockchain technology to create a stable, cryptographically secured coin similar to fiat currency that will reduce volatility for investors in the cryptocurrency market. Stablecoins are, therefore, particularly useful for those investors who want to redeem or exit their positions in the market. In order to peg a stablecoin to a fiat currency, the coin can either be backed by cash-equivalent reserves such as Treasury bills or backed by a smart contract on the blockchain. The smart contract ensures that the peg holds by buying or selling the required number of coins once preset conditions are met.[6]

Figures 3 and 4 show that while the tether coin experienced noticeable volatility, it has nevertheless remained stable over recent years. While on average the value of the Tether is highly correlated with the value of the U.S. dollar, the volume is highly dependent on Bitcoin. The correlation between the Tether volume and the Bitcoin volume is about 91%. This correlation suggests that the observed volatility in Bitcoin has real implications for stablecoins. During periods of significant volatility in Bitcoin, redemption risks in stablecoins are likely to increase, leading to potential rollover risk in the cryptocurrency market.[7]

[5] The updated BSV 1.0.7 (released, 2021) has no block size limit and the protocol can handle scalable transactions but it is not yet clear whether this will lead to scalability at the global level (MNP, 2021).

[6] Financial intermediaries can also issue stablecoins. J. P. Morgan uses "JPM Coin" for intra-day repurchase agreements and for liquidity management. But there is still ongoing debate as to whether these types of coins are actually stable coins or digitized alternative forms of payment system (Morgan, 2020).

[7] See: Liao and Caramichael (2022) Gorton and Zhang (2021) for some discussion on stablecoins.

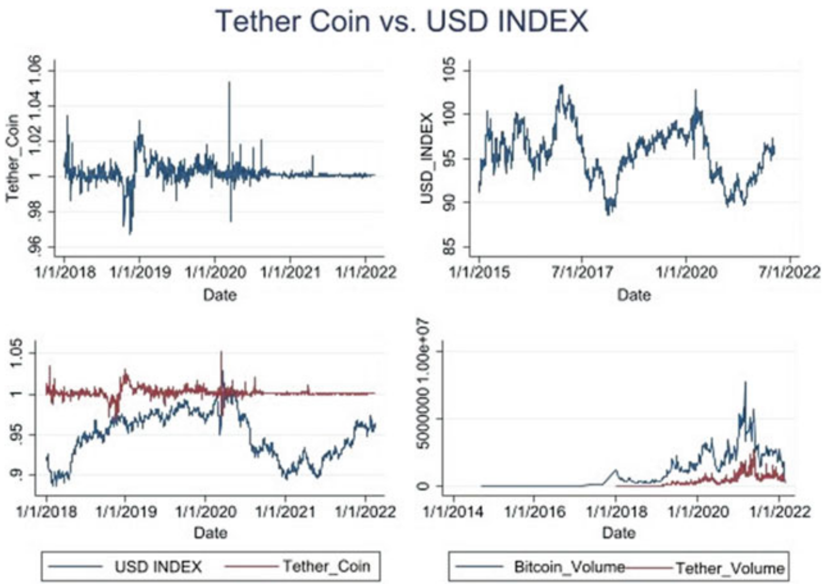


Fig. 3 Tether coin versus U.S. dollar (USD) index

2.2 Bitcoin vs Gold: A Store of Value?

Can Bitcoin replace gold and other precious metals as a store of value and a hedge against inflation? First, for an asset to be considered a store of value, it must meet several requirements: (1) Maintain purchasing power over a long duration of time (2) The asset must be easy to transport and durable, and (3) The asset should have some element of inherent value, either due to historical adoption, government backing or because of economic and industrial use. Gold meets all of these three characteristics. Gold is limited in supply and, as a result, tends to maintain its purchasing power over time, making it a reliable hedge against inflation (Capie et al., 2005). Gold is valued for its aesthetic qualities and does not degrade over time. Gold has also historically been accepted as a store of value (Graeber, 2012; Taleb, 2021; Sargent and Wallace, 1983) and, as a result, provides some protection against regional and national political uncertainties. Gold-backed exchange-traded funds make it easy to transport, trade, and own gold as a store of wealth, and investors do not necessarily need to hold the actual physical gold. Bitcoin is limited in

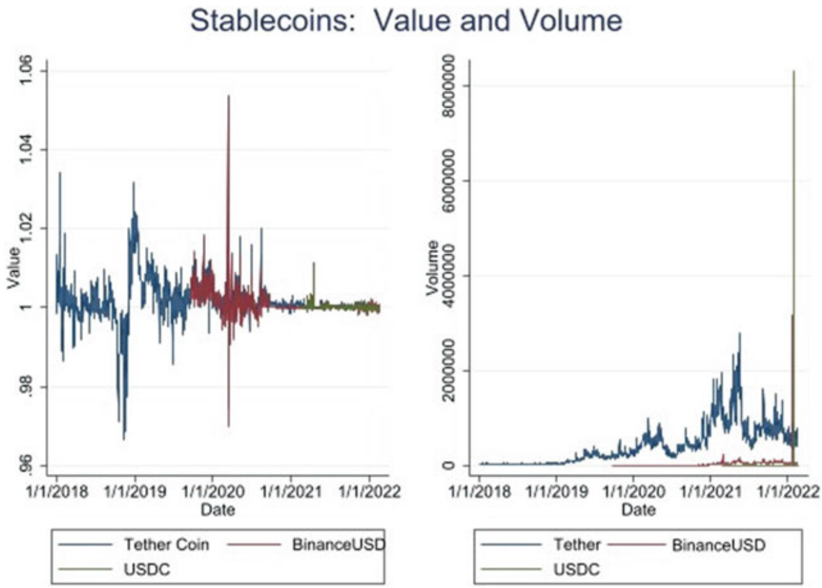


Fig. 4 Value and volume of Stablecoins

supply since there are only 21 million hard-coded coins, and the supply growth is expected to decrease over time due to the deflationary nature of the Bitcoin protocol (Lipton, 2021; Nakamoto, 2008). This supply limit suggests that the value of Bitcoin cannot be devalued by any central authority such as a Central Bank. However, Bitcoin has yet to be globally accepted as a store of wealth, partially because of its high volatility and because it has only existed for a decade. Therefore, the volatility of Bitcoin significantly weakens its ability to be an effective store of value and a diversifier in an individual's portfolio.

Additionally, during the early nineteenth century, gold historically served as an automatic stabilizing mechanism. Most major currencies were backed by or linked to gold, and as a result, gold has historically served as an important asset during market downturns. In particular, investors tend to hold gold when a currency depreciates in value and reduce their gold holdings when a currency appreciates. In this regard, gold has served as an effective exchange rate hedge, both against the decline in domestic currency's purchasing power and against domestic currency's purchasing power relative to a basket of foreign currencies (Capie et al., 2005). Unlike fiat currency or gold, the demand for Bitcoin is unpredictable and difficult to stabilize as price appreciation encourages hoarding, which could lead to deflation if Bitcoin is the base currency in an economy (Dowd, 2014; Selgin, 2014, Bohme et al., 2015. Thus, it is unclear yet what role Bitcoin would play during periods of significant currency fluctuations and whether Bitcoin can serve as an effective exchange rate hedge. Additionally, Fig. 5 shows that Bitcoin performed poorly as an inflation hedge relative to gold during a market downturn in March 2020.

Unlike gold, Bitcoin is not a homogeneous asset as there exists a continuous stream of competing cryptocurrencies assets, making Bitcoin less suitable as an inflation or a currency hedge for investors. Bitcoin also has no obvious industrial usage. If anything, the cost of Bitcoin mining and its energy consumption is significantly higher than the cost of minting fiat currency (Antonopoulos & Wood, 2018). Figure 6 shows that mining difficulty has been steadily increasing, while block size, the

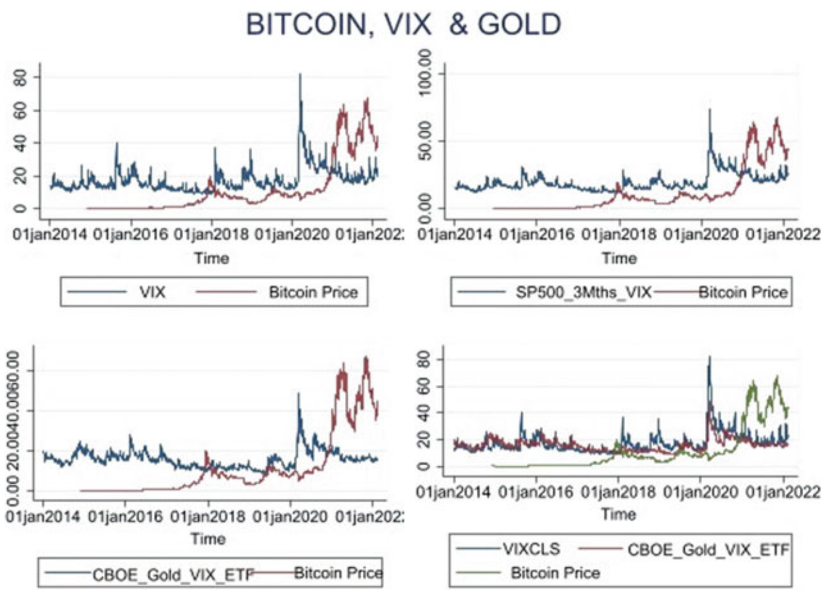


Fig. 5 Comparing the price of Bitcoin, the volatility index (VIX), and the price of gold

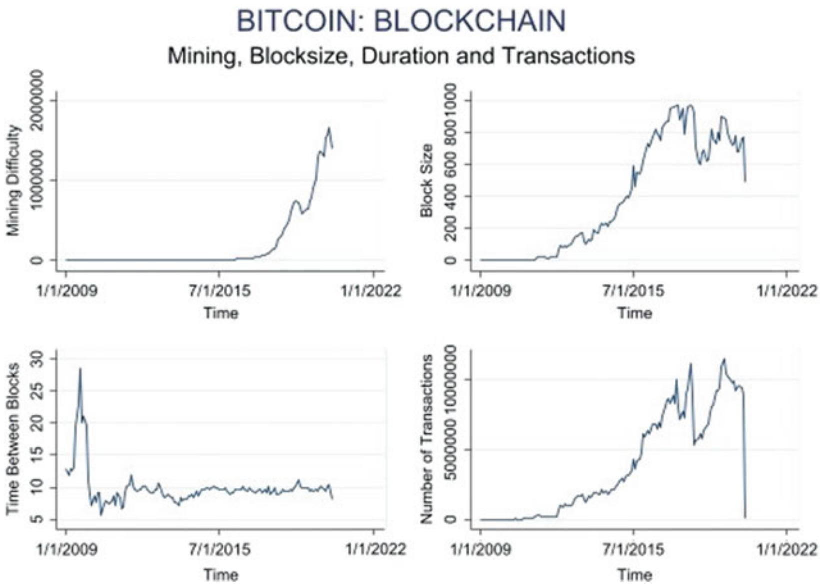


Fig. 6 Blockchain technology mining, blocksize, duration, and transactions time series

time between blocks, and the number of transactions have declined. In 2019, the average transaction in Bitcoin consumed about 0.51 megawatts hours, and Bitcoin protocol energy consumption was about 0.3% of global energy consumption (Lipton, 2021; Feenan et al., 2021). Note that the cost of gold mining, processing, production, and energy consumption is generally already priced in due to its long history of usage. This priced-in mechanism suggests that in order for Bitcoin to replace gold in the near future, its energy consumption cost has to decline significantly. Otherwise, it is not an effective alternative to gold at this stage, and its diversification benefits during a market downturn remain questionable.

2.3 Bitcoin: Investment and Diversification Role

Is Bitcoin an investment? If so, where does it fit in an individual's portfolio? Investors must be able to value an asset to determine its relative impact on their portfolio. An asset is likely to have a diversification role if it is positively correlated with another asset in the portfolio, and it has a hedging role if it is negatively correlated with an asset in the portfolio (Baur & Lucey, 2010; Chan et al., 2019). Bitcoin has no fundamentals, and therefore it is difficult to value.[8] Since Bitcoin has no intrinsic value or industrial usage, its price can range from zero to infinity. The price can be zero because Bitcoin neither pays out dividends nor has future earnings; therefore, the present value of Bitcoin's price is zero (Taleb, 2021). Moreover, the price of Bitcoin can rise to infinity due to irrational exuberance. In particular, Bitcoin prices are driven mainly by market sentiments and price appreciation (Weber, 2016). The expectation of a continuous increase in prices divorced from any fundamental value can lead to an irrational bubble (Dale et al., 2005; Shiller, 2015). The fluctuations in the price of Bitcoin, as shown in Figs. 1 and 2, provide some opportunities for speculative trading (Cheah & Fry, 2015; Dwyer, 2015; Blanchard & Watson, 1982).

So why is there institutional interest in Bitcoin? First, the price fluctuations in Bitcoin and other cryptocurrencies provide some alpha and profit-making opportunities. Given that other cryptocurrencies are currently priced in Bitcoin, it also provides some arbitrage opportunities. Arbitrage opportunities exist because of price differentials between crypto-linked assets in traditional finance and on-chain products, making Bitcoin potentially valuable in portfolio management (Dyhrberg, 2016; Karniol-Tambour et al., 2022; Makarov & Schoar, 2020; Tully & Lucey, 2017, Briere et al., 2015). Secondly, institutional investors might treat Bitcoin as a long-duration asset and anticipate that there would be future opportunities to offload at a higher price due to its limited supply and potential price appreciation. Thirdly, institutional investors are investing indirectly in Bitcoin and the cryptocurrency space via venture capitals that use blockchain technology, as it aligns well with their investment mandates (Karniol-Tambour et al., 2022, Bouri et al., 2017). In particular, high-frequency trading (HFT) and long-short equity funds that use a cash-and-carry strategy have netted an average return of at least 10% by buying Bitcoin and selling CME futures. Institutional investors are

[8] Theoretical, the value of Bitcoin (as an asset) is approximately the discounted sum of its cash flows, service flow, and some speculative or heterogeneous beliefs regarding the asset.

therefore able to reduce risk exposure from investing in the cryptocurrency space by either investing in the early stages of these exchanges or by using sophisticated trading.[9]

One advantage of investing in Bitcoin is that it provides some protection against inflation as the central bank cannot devalue it. However, this protection comes at the expense of increasing volatility in the portfolio. These speculations in the cryptocurrency space suggest that individuals should be concerned about the level of exposure in their portfolios. Profit-making opportunities for retail traders are likely to decline as institutional investors and hedge funds using sophisticated trading strategies take advantage of the mispricing and other market inefficiencies in the cryptocurrency space. The fact that the hedging and diversification abilities of Bitcoin depend on data frequency,[10]in the long run, can only make Bitcoin less desirable relative to alternative assets. Figures 7a,b show that Bitcoin returns are more volatile than Standard and Poor's 500 (S&P 500) returns over the same duration. Trades per minute have also risen across all exchanges, as can be seen in Fig. 8. Furthermore, liquidity as a proxy of the bid-ask spread, as seen in Fig. 9, has also been steadily increasing, reflecting a growing interest in Bitcoin.

Additionally, the price of Bitcoin remains high, liquidity is low relative to major indices, and there is some evidence of price manipulation in the cryptocurrency space (Griffin & Shams, 2020). These factors could potentially limit and discourage ownership of Bitcoin and related cryptocurrencies. However, the rise of exchange-traded funds (ETFs) in this space provides some opportunities for small and regular investors to have an indirect exposure to the cryptocurrency market.

2.3.1 Bitcoin: Political Uncertainty and Dictatorial Regimes

Is Bitcoin a safe haven? Bitcoin can provide a channel for transferring large funds across international borders independent of any third party or entity. This non-centralization of Bitcoin provides some protection against dictatorial regimes or during periods of political uncertainty. Case in point, Fig. 10 shows that Bitcoin's price and returns increased significantly

[9] Some of these strategies include: tail-risk hedging, factor-based investing, stock-picking, and asymmetric bets using options that leverage the inefficiencies in the crypto-market.

[10] See: Bouri et al. (2017) ;Chan et al. (2019); Stavroyiannis, (2018).

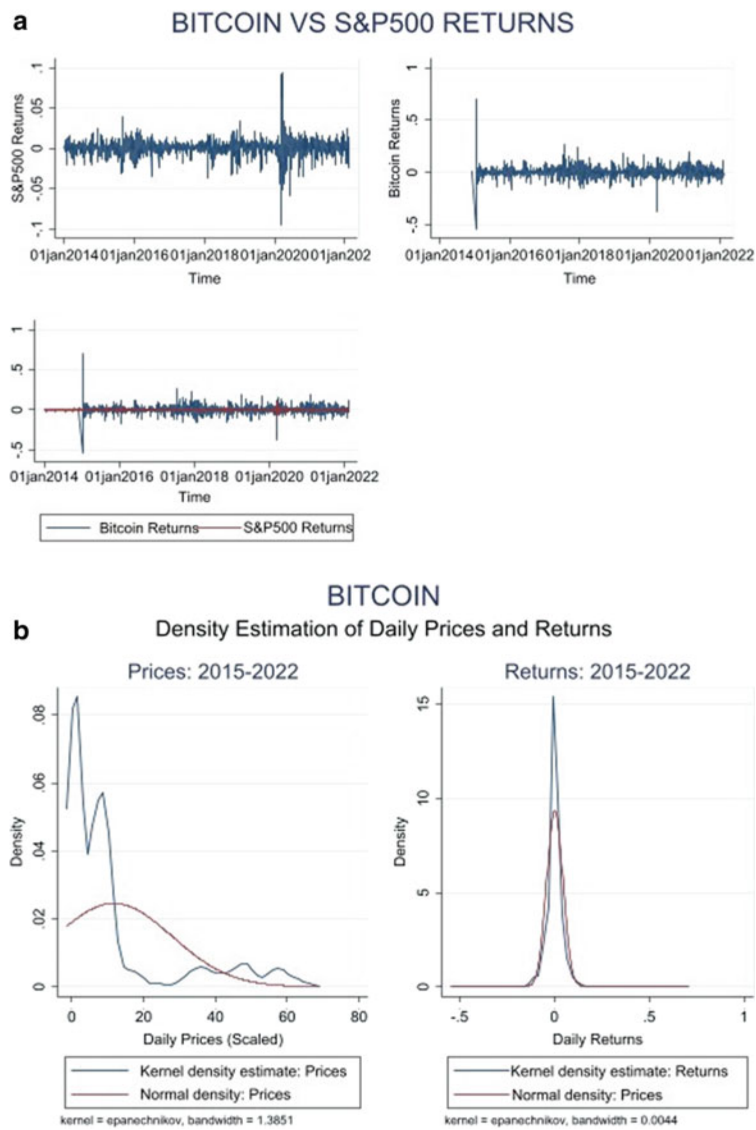


Fig. 7 **a** Bitcoin versus Standard and Poor's 500 (S&P500) returns. **b** Kernel density of daily Bitcoin prices and returns

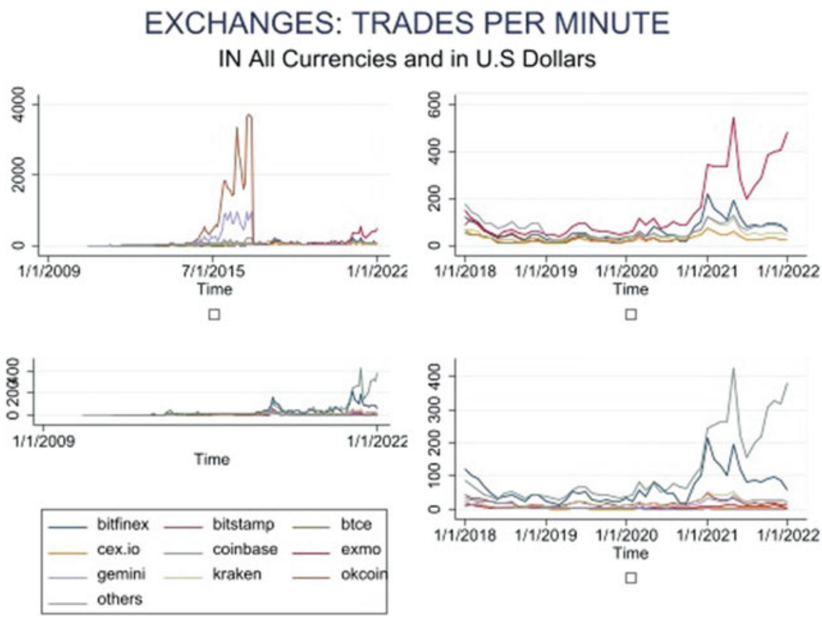


Fig. 8 Trades per minute on cryptocurrency exchanges

compared to gold during the immediate onset of the ongoing Russian–Ukrainian conflict. This result illustrates the potential that Bitcoin can serve as a safe asset ("flight-to-quality") during periods of significant political uncertainty.

2.4 Is Bitcoin a Collectible Asset?

Collectibles are a form of alternative investment. This form of investment generally includes fine arts, baseball cards, rare coins, comic books, and rare books. In addition to the pecuniary benefits, alternative investments generally provide some subjective utility to the owners. Bitcoin can be considered a "rare" collectible asset since there are only 21 million hard-coded Bitcoin, and 90% have already been mined; thus, it has a residual value that makes it valuable to hold into the future. The holders of Bitcoin might also infer some value from both the embedded technology and in being a part of a new and potentially useful innovative idea. The Bitcoin

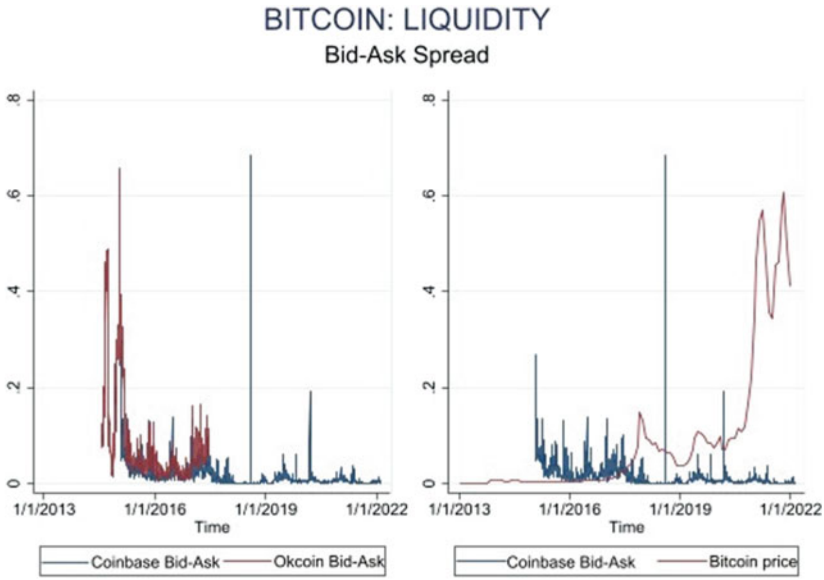


Fig. 9 The liquidity of Bitcoin

protocol has become a useful baseline for the new wave of cryptocurrencies, building smart contract-based tokens and other distributed ledger technology (DLT). Smart contracts are simply a set of rules stored in the blockchain that are automatically executed once the set conditions are met, thereby facilitating exchanges and transactions independent of a third party. The utility that comes from being at the forefront of this new movement and in leveraging big data in the cryptocurrency space makes Bitcoin a valuable collectible asset.

Collectibles are transferable inter-generational assets, and given that the block creation fee is projected to go down to zero in the year 2140, this could potentially explain why about 70% of Bitcoins are contained in less active and dormant accounts (Cheah & Fry, 2015, Chuen, 2015). Additionally, the rise of non-fungible tokens (NFTs), which are digital collectibles that enable users to authenticate ownership as transactions recorded on a blockchain, demonstrates that Bitcoin and other cryptocurrencies have some features in common with other collectible assets.

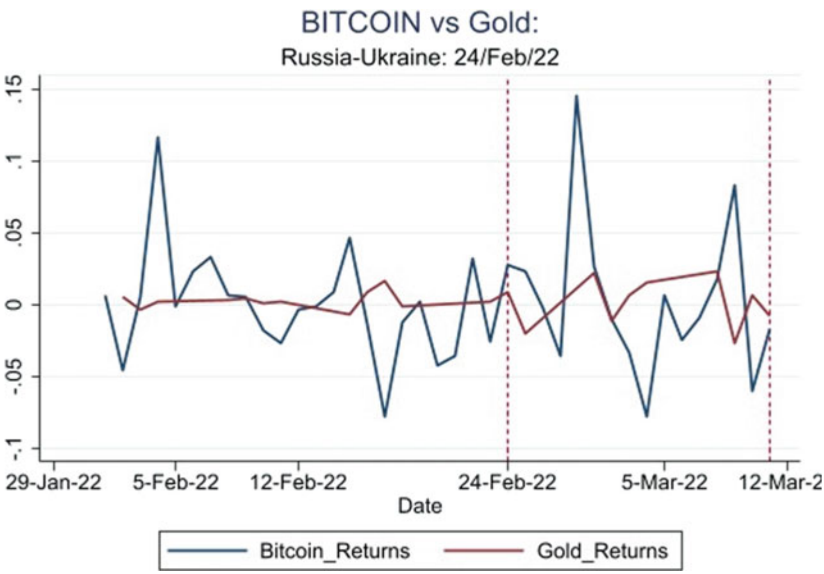


Fig. 10 Comparing the returns for Bitcoin and gold

Unlike Bitcoin and other cryptocurrencies, NFTs have an intrinsic value since they cannot be exchanged like-for-like. The intrinsic value of NFTs increases their applicability and marketability to a wider audience. NFTs employ Blockchain technology in two major ways. First, blockchain technology is used to create play-to-earn games in which users are incentivized to play the game via a reward, usually in the form of a token. Second, blockchain technology is used to create a metaverse, a virtual world in which various tokens can be traded for different assets, including virtual properties and artworks (Aharon & Demir, 2021).

3 Discussion

3.1 *What is Bitcoin's Real Contribution: Cryptocurrencies, Big Data, and Blockchain Technology*

What is the long-term value and contribution of Bitcoin to society? The Bitcoin protocol and blockchain technology have created many new cryptocurrency assets and digital products. For example, Ethereum utilizes Bitcoin's protocol to build a decentralized transaction-based state machine that uses a cryptographic hash to collate transactions into a blockchain (Wood, 2014; Okamoto & Ohta, 1992. Thus, the built-in Turing-programming language in the Ethereum blockchain can quickly create smart contracts from an arbitrary set of codes (Buterin, 2013; Lipton, 2021, Bhutoria, 2020) and the PoW simply then ensures absolute confidence in the future viability of the protocol since each mined block has a reward attached to it. Additionally, Ethereum provides a potential solution to the application-specific integrated circuits (ASICs) problem in the Bitcoin protocol via the Ethash algorithm (Buterin, 2013; Poon and Dryja, 2016; Jentsch, 2015). However, Ethereum has scalability problems since supply is limited to only 18 million ETH per year. Ethereum's underlying protocol is quite costly to use, as smart contracts tend to require a large amount of collateral in their operations (Antonopoulos & Wood, 2018; Lipton, 2021; Lipton & Hardjono, 2021; Lipton et al., 2016). Bitcoin's blockchain technology is also currently being utilized to build decentralized finance (DeFi). DeFI is based on a Consensus as a Service (CaaS) and can be used to create smart contracts-based (cryptocurrency) exchanges independent of a third party (Lipton, 2021).

Furthermore, blockchain technology is a potentially useful tool in solving big data challenges. For example, healthcare providers are faced with challenges ranging from access to patients' health data, legal issues, secure storage, and ownership of the data. Blockchain technology can provide a secure structure in which healthcare providers can store the metadata of patients in a blockchain and then provide the patient with a unique key that can be used to access their health data anywhere (Gupta et al., 2016; Rapke, 2016). Applications such as "Storj" that use blockchain technology to ensure secured peer-to-peer authentication of storage contracts are potential solutions to big data challenges in the healthcare sector (Zhang et al., 2019). Storj uses smart contracts to manage, record, and keep a timestamp of data sharing.

Blockchain technology can also be used to protect intellectual property rights and authenticate ownership of digital art. An interesting implementation and application of blockchain technology in this area is the "Secure Public Online Ownership Ledger" (SPOOL), which can be used for documenting transactions, transferring ownership of each edition of the artwork, and recording it in a blockchain, which allows for tracking and authentication of ownership (Dejonghe &

McConaughy, 2016; McConaghy et al., 2016; Karafiloski & Mishev, 2017; McConaghy & Holtzman, 2015). Some applications of blockchain technology in marketing and supply chain management include "Omnilytics" and "Provenance" (Deepa et al., 2022). Blockchain in these applications is used to collate sales, marketing, industry trends, and product information data during each point in the supply chain (Kim & Laskowski, 2018). For example, Walmart and IBM have utilized blockchain technology (Hyperledger Fabric) to create a food traceability system, and early tests of the system have shown that blockchain technology can significantly reduce the time it takes to trace the provenance of produce in the supply chain from days to seconds (Hyperledger, 2020). Blockchain-based applications such as "Rubix" provide a decentralized trading platform where users can buy and sell cryptocurrencies and digital assets independent of a financial intermediary.

Blockchain technology is also useful in addressing some of the big data challenges in the financial services sector. Because data management is critical for financial institutions, most transactions generally incur some fees. These fees and charges erode returns to banks' clients and shareholders. Blockchain technology can ensure that banks monitor, detect and prevent fraudulent transactions at a minimal cost. Signature Bank has launched a blockchain-based payment platform called "Signet." The platform provides a channel via which Signature bank's commercial clients can transact with other commercial clients at zero cost, effectively eliminating the need for a third party. Blockchain technology has also been used to address some of the challenges in the securities lending markets. For example, Deutsche Boerse launched a distributed ledger technology in swap trading, which has reduced the cost of trading in these types of financial instruments (Morgan, 2020). These applications demonstrate that blockchain technology goes beyond simply the creation and minting of new digital coins and can potentially provide solutions to real challenges for individuals and businesses. These applications also demonstrate that the emerging technologies underpinned by blockchain or smart contracts can be engines for economic growth.

3.2 Government Regulations

Understanding the role of government regulations in the cryptocurrency space is essential as regulations affect both Bitcoin's long-term adoption as a currency and tax treatment as an asset. Aristotle argued that money derives its value not from nature but from the law and can therefore be altered or abolished at will (Crisp, 2014). This idea that Aristotle presented clearly demonstrates that retail and institutional investors' potential global adoption of Bitcoin largely depends on governmental regulations.

Why should the government be interested in the cryptocurrency space? And why should investors care about regulations? There are two critical reasons. First, Bitcoin is a potential money laundering channel, which could impact the value of the reserve currency (U.S. dollar) and other major currencies. If the U.S. government decides to ban Bitcoin and related digital currencies, this would automatically drive their values to zero, making them less desirable for investors. If the government decides to introduce its own digital currency alongside Bitcoin, then this can only increase Bitcoin's price volatility and weaken its diversification role. Indeed, the Chinese government's ban on cryptocurrency mining and initial coin offerings in 2017 led to a precipitous drop of about 7.8% in Bitcoin prices. China, driven by concerns regarding the potential impact of a decentralized digital currency on monetary policy and the subsequent impact on fiat currency (Renminbi), is in the process of introducing a Central Bank Digital Currency (CBDC). Cryptocurrency users are more likely to use CBDC than their alternative decentralized digital currencies.[11]

Secondly, Bitcoin and other cryptocurrencies are potential sources of revenue since they can be treated as taxable investment vehicles. Some countries, such as Canada, consider cryptocurrencies as commodities, and these are taxed as either business income or as capital gains (50%). If a cryptocurrency is used to exchange goods and services, this is treated by the Canadian government as a barter transaction. Additionally, Hungary taxes any cryptocurrency income at 15% once it has been converted to fiat currency regardless of the source(s). The United States government's Internal Revenue Service (IRS), as of the 2022 tax year, treats Bitcoin and other cryptocurrencies as "property" and, therefore, as taxable assets. The long-term impact of this IRS policy is not yet clear, but it can only facilitate wider adoption and lead to a further increase in the price volatility of Bitcoin. Taxes will further erode some of the gains, making Bitcoin less

[11] Note that following China's ban, some miners simply moved their rigs to energy-rich countries such as Kazakhstan (Oxford Analytica, 2021).

attractive to investors relative to alternative assets that might have more favorable tax treatment.

4 Concluding Thoughts

Based on the findings presented in this chapter, we determine that Bitcoin is neither gold nor currency but a tradable asset and an alternative form of investment. Bitcoin also exhibits some features as an investment asset that are similar to collectibles. The true value of Bitcoin lies not in its speculative nature (price appreciation) but in the embedded technology (blockchain, DeFi, and Distributed Ledger Technologies), which has the long-term potential to revolutionize traditional finance. Blockchain technology can solve big data challenges in collecting, organizing, controlling, and storing a large amount of data. Bitcoin's long-term survivability and viability as an asset will largely depend on its diversification role, tax treatment, and government regulations.

References

Aharon, D. Y., & Demir, E. (2021). NFTs and asset class spillovers: Lessons from the period around the COVID-19 pandemic. *Finance Research Letters, 10*, 2515. https://doi.org/10.1016/j.frl.2021.102515

Antonopoulos, A. M., & Wood, G. (2018). *Mastering ethereum* (1st Ed.). O'reilly Media. https://www.oreilly.com/library/view/masteringethereum/9781491971932/

Akcora, C. G., Dey, A. K., Gel, Y. R., & Kantarcioglu, M. (2018) Forecasting bitcoin price with graph chainlets. In: D. Phung, V. Tseng, G. Webb, B. Ho, M. Ganji, & L. Rashidi (Eds.), *Advances in knowledge discovery and data mining*. PAKDD 2018. Lecture Notes in Computer Science (1st Ed., pp. 765–776). Springer. https://doi.org/10.1007/978-3-319-93040-460

Almudhaf, F. (2018). Pricing efficiency of bitcoin trusts. *Applied Economics Letters, 25*(7), 504–508. https://doi.org/10.1080/13504851.2017.1340564

Baur, D. G., & Lucey, B. M. (2010). Is gold a hedge or a safe haven? An analysis of stocks, bonds and gold. *Financial Review, 45*(2), 217–229. https://doi.org/10.1111/j.1540-6288.2010.00244.x

Bhutoria, R. (2020). Addressing persistent bitcoin criticisms. *Fidelity Digital Assets*. https://www.fidelitydigitalassets.com/articles/addressing-bitcoin-criticisms

Blanchard, O. J., & Watson, M. (1982). Bubbles, rational expectations and financial markets. In P. Wachtel (Eds.), *Crises in the economic and financial structure* (1st Ed., pp. 295–316). D.C. Heath and Company. https://doi.org/10.3386/w0945

Bohme, R., Christin, N., Edelman, B., & Moore, T. (2015). Bitcoin: Economics, technology, and governance. *Journal of Economic Perspectives, 29*(2), 213–238. https://doi.org/10.1257/jep.29.2.213

Bouri, E., Georges, M., David, A., & Hagfors, R. (2017). On the hedge and safe haven properties of Bitcoin: Is it really more than a diversifier? *Finance Research Letters 20*, 192–198, S1544612316301817. https://doi.org/10.1016/j.frl.2016.09.025

Briere, M., Oosterlinck, K., & Szafarz, A. (2015). Virtual currency, tangible return: Portfolio diversification with bitcoin. *Journal of Asset Management, 16*(6), 365–373.

Buterin, V. (2013). *Ethereum white paper*. https://github.com/ethereum/wiki/wiki/White-Paper

Capie, F., Mills, T. C., & Wood, G. (2005). Gold as a hedge against the dollar. *Journal of International Financial Markets, Institutions and Money, 15*(4), 343–352. https://doi.org/10.1016/j.intfin.2004.07.002

Chan, W. H., Le, M., & Wu, Y. W. (2019). Holding bitcoin longer: The dynamic hedging abilities of bitcoin. *The Quarterly Review of Economics and Finance, 71*, 107–113. https://doi.org/10.1016/j.qref.2018.07.004

Cheah, E. T., & Fry, J. (2015). Speculative bubbles in bitcoin markets? An empirical investigation into the fundamental value of Bitcoin. *Economics Letters, 130*, 32–36. https://doi.org/10.1016/j.econlet.2015.02.029

Chuen, D. L. (2015). *Handbook of digital currency: Bitcoin, innovation, financial instruments, and big data* (1st Ed.). Academic Press. https://doi.org/10.1016/C2014-0-01905-3

Crisp, R. (2014). Aristotle: Nicomachean ethics. *Cambridge University Press*. https://doi.org/10.1017/CBO9781139600514

Dale, R. S., Johnson, J. E., & Tang, L. (2005). Financial markets can go mad: Evidence of irrational behaviour during the South Sea Bubble 1. *The Economic History Review, 58*(2), 233–271. https://doi.org/10.1111/j.1468-0289.2005.00304.x

Davies, G. (2010). *History of money*. University of Wales Press.

Deepa, N., Pham, Q. V., Nguyen, D. C., Bhattacharya, S., Prabadevi, B., Gadekallu, T. R., Maddikunta, P. K. R., Fang, F., & Pathirana, P. N. (2022). A survey on blockchain for big data: Approaches opportunities and future directions. *Future Generation Computer Systems, 131*, 209–226. S0167739X22000243. https://doi.org/10.1016/j.future.2022.01.017

DeJonghe, D., & McConaughy, T. (2016). *SPOOL protocol*. https://github.com/ascribe/spool

Dowd, K. (2014). New private monies: A bit-part player? *Institute of Economic Affairs Monographs*, Hobart Paper.
Dwyer, G. P. (2015). The economics of Bitcoin and similar private digital currencies. *Journal of Financial Stability, 17*, 81–91. https://doi.org/10.1016/j.jfs.2014.11.006
Dyhrberg, A. H. (2016). Bitcoin gold and the dollar—A GARCH volatility analysis. *Finance Research Letters, 16*, 85–92. S1544612315001038. https://dopi.org/10.1016/j.frl.2015.10.008
Easley, D., O'Hara, M., & Basu, S. (2019). From mining to markets: The evolution of bitcoin transaction fees. *Journal of Financial Economics, 134*(1), 91–109. https://doi.org/10.1016/j.jfineco.2019.03.004
Feenan, S., Heller, D., Lipton, A., Morini, M., Ram, R., Sams, R., & Barrero Zalles, D. (2021). Decentralized financial market infrastructures: Evolution from intermediated structures to decentralized structures for financial agreements. *The Journal of FinTech, 1*(2). https://doi.org/10.1142/S2705109921500024
Foley, S., Karlsen, J. R., & Putniņš, T. J. (2019). Sex, drugs, and bitcoin: How much illegal activity is financed through cryptocurrencies? *The Review of Financial Studies, 32*(5), 1798–1853. https://doi.org/10.1093/rfs/hhz015
Goetzmann, W., & Goetzmann, W. N. (2017). Money changes everything. *Princeton University Press*. https://doi.org/10.1515/9781400888719
Gorton, G. B., & Zhang, J. (2021). Taming wildcat stablecoins. *University of Chicago Law Review, 90*, Forthcoming. https://doi.org/10.2139/ssrn.3888752
Graeber, D. (2012). *Debt: The first 5000 years*. Penguin. https://tinyurl.com/3hx4fnae
Griffin, J. M., & A. Shams. (2020). Is bitcoin really untethered? *The Journal of Finance, 75*(4), 1913–1964. https://doi.org/10.1111/jofi.12903
Gupta, N., Jha, A., & Ro, S. (2016). *Adopting blockchain technology for electronic health record interoperability*. Cognizant Technology Solutions.
Hyperledger. (2020). *How Walmart brought unprecedented transparency to the food supply chain with hyperledeger fabric*. https://tinyurl.com/e38rhxys
Jentsch, M. (2015). Ubiquitous annotation visualization-concept and rapid prototyping framework. *Doctoral dissertation, Dissertation, Aachen, Techn. Hochsch*. https://core.ac.uk/download/pdf/36622081.pdf
Karafiloski, E., & Mishev, A. (2017). Blockchain solutions for big data challenges: A literature review. In IEEE EUROCON 2017-17th International Conference on SmartTechnologies (pp.763–768). IEEE. https://doi.org/10.1109/EUROCON.2017.8011213
Karniol-Tambour, R., Tan, D., & Tsarapkina, D. (2022). *The evolution of institutional exposure to cryptocurrencies and blockchain technologies*. Bridgewater Inc. https://tinyurl.com/2p8856xn

Keynes, J. M. (2018). The general theory of employment, interest, and money. *Springer*. https://doi.org/10.1007/978-3-319-70344-21

Kim, H. M., & Laskowski, M. (2018). Toward an ontology-driven blockchain design for supply-chain provenance. *Intelligent Systems in Accounting, Finance and Management, 25*(1), 18–27. https://doi.org/10.1002/isaf.1424

Liao, G. Y., & Caramichael, J. (2022). Stablecoins: Growth potential and impact on banking. *International finance discussion papers, (1334)*, 1–26. https://doi.org/10.17016/IFDP.2022.1334

Lipton, A. (2021). Cryptocurrencies change everything. *Quantitative Finance, 21*(8), 1257–1262. https://doi.org/10.1080/14697688.2021.1944490

Lipton, A., & Treccani, A. (2021). Blockchain and distributed ledgers: Mathematics, technology, and economics. *World Scientific*. https://doi.org/10.1142/11857

Lipton, A., & Hardjono, T. (2021). Blockchain intra-and interoperability. In V. Abich, J. Birge & G. Hilary (Eds.), *Innovative technology at the interface of finance and operations. Springer Series in Supply Chain Management*, forthcoming, Springer Nature.

Lipton, A., Shrier, D., & Pentland, A. (2016). *Digital banking manifesto: the end of banks?* Massachusetts Institute of Technology. https://tinyurl.com/3ubp5ahu

Makarov, I., & Schoar, A. (2020). Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics, 135*(2), 293–319. https://doi.org/10.1016/j.jfineco.2019.07.001

McConaghy, T., & Holtzman, D. (2015). Towards an ownership layer for the internet. *Ascribe*. https://tinyurl.com/2p94htvr

McConaghy, T., Marques, R., Müller, A., De Jonghe, D., McConaghy, T., McMullen, G., Henderson, R., Bellemare, S., & Granzotto, A. (2016). *Bigchaindb: a scalable blockchain database*. White paper, BigChainDB. https://tinyurl.com/5y3rptp2

McLeay, M., Radia, A., & Thomas, R. (2014). Money creation in the modern economy. *Bank of England Quarterly Bulletin*. https://ssrn.com/abstract=2416234

MNP. (2021, August 25). *The original Bitcoin protocol: What is it and why does it matter?* https://tinyurl.com/yu5mrrsp

Morgan, J. P. (2020, February 21). *Blockchain, digital currency and cryptocurrency: Moving into the mainstream?* (J.P. Morgan. Global Research Reports). https://www.jpmorgan.com/insights/research/reports

Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review, 21*, 260. https://www.debr.io/

Okamoto, T., & Ohta K. (1992). Universal electronic cash. In: J. Feigenbaum (Eds.), *Advances in cryptology—CRYPTO '91. CRYPTO 1991. Lecture Notes*

in Computer Science (Vol. 576). Springer. https://doi.org/10.1007/3-540-46766-1_27
Oxford Analytica. (2021). Kazakh law plays catch-up with cryptocurrency miners. *Expert Briefings*. https://doi.org/10.1108/OXAN-ES262378
Poon, J., & Dryja, T. (2016). *The bitcoin lightning network: Scalable off-chain instant payments*. https://tinyurl.com/5fwadhjn
Rapke, T. (2016). *Blockchain technology & the potential for its use in healthcare*. https://tinyurl.com/4fk5n5ts
Sargent, T. J., & Wallace, M. (1983). A model of commodity money. *Journal of Monetary Economics, 12*(1), 163–187. https://doi.org/10.1016/0304-3932(83)90055-7
Selgin, G. (2014). Bitcoin: Problems and prospects. *Hillsdale University's 2014 Free Market Forum*. https://tinyurl.com/mryphx6p
Shiller, R. J. (2015). *Irrational exuberance*. Princeton University Press. https://tinyurl.com/5ycfr48b
Smithin, J. (2002). What is money? *Routledge*. https://doi.org/10.4324/9780203072691
Song, Y. D., & Aste, T. (2020). The cost of Bitcoin mining has never really increased. *Frontiers in Blockchain*. https://doi.org/10.3389/fbloc.2020.565497
Stavroyiannis, S. (2018). Value-at-risk and related measures for the bitcoin. *The Journal of Risk Finance, 19*(2), 127–136. https://doi.org/10.1108/JRF-07-2017-0115
Stoll, C., Klaaßen, L., & Gallersdörfer, U. (2019). The carbon footprint of bitcoin. *Joule, 3*(7), 1647–1661. https://doi.org/10.1016/j.joule.2019.05.012
Taleb, N. (2021). Bitcoin, currencies, and fragility. *Quantitative Finance, 21*(8), 1249–1255. https://doi.org/10.1080/14697688.2021.1952702
Thum, M. (2018). The economic cost of bitcoin mining. In *CESifo Forum* (Vol. 19, No. 1, pp. 43–45). München: ifo Institut-Leibniz-Institut für Wirtschaftsforschung an der Universität München. http://hdl.handle.net/10419/181201
Tully, E., & Lucey, B. M. (2007). A power GARCH examination of the gold market. *Research in International Business and Finance, 21*(2), 316–325. S0275531906000353. https://doi.org/10.1016/j.ribaf.2006.07.001
Weber, B. (2016). Bitcoin and the legitimacy crisis of money. *Cambridge Journal of Economics, 40*(1), 17–41. https://doi.org/10.1093/cje/beu067
Wood, G. (2014). Ethereum: A secure decentralized generalized transaction ledger. *Ethereum Project Yellow Paper, 151*, 1–32. https://ethereum.github.io/yellowpaper/paper.pdf

Yermack, D. (2015). Is Bitcoin a real currency? An economic appraisal. In *Handbook of digital currency* (1st Ed., pp. 31–43). Academic Press. https://doi.org/10.1016/B978-0-12-802117-0.00002-3

Zhang, X., Grannis, J., Baggili, I., & Beebe, N. L. (2019). Frameup: An incriminatory attack on Storj: A peer to peer blockchain enabled distributed storage system. *Digital Investigation, 29*, 28–42, S1742287618303438. https://doi.org/10.1016/j.diin.2019.02.003

Culture, Digital Assets, and the Economy: A Trans-National Perspective

John Fan Zhang, Zehuang Xu, Yi Peng, Wujin Yang, and Haorou Zhao

1 Introduction

In society today, the development of digital assets has dramatically improved our way of life and means of production and consumption. The use of digital assets refers mainly to digital payments including mobile payments, paying bills through the internet, and online shopping. The COVID-19 pandemic, for example, has restricted offline consumption and promoted new habits such as online shopping, online education, and online film and television consumption. During the pandemic period, the penetration rate of online shopping rose, and the population accessing it gradually shifted from the young to the middle-aged and elderly. Furthermore, digital assets have a dramatic impact not only on people's consumption but also on investments. There is a trend for more and more people to choose online investment and financial management, such as buying bonds and funds through specific software applications, as opposed to choosing traditional in-person banking. As a result, the combination of technology, the internet, and finance is breaking new

J. F. Zhang (✉) · Z. Xu · Y. Peng · W. Yang · H. Zhao
Macau University of Science and Technology, Taipa, Macao
e-mail: fanzhang@must.edu.mo

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_9

ground and creating more possibilities for the development of the financial industry. As Osei and Kim (2020) and Xue (2020) report, this development has the potential to benefit the economy.

International scholars have been involved in intense discussions about how internet developments are impacting global economic growth. Jorgenson et al. (2008) claim that Information and Communication Technology (ICT) is a precondition of national economic growth in the modern world. Oliner and Sichel (2000) state that productivity growth in the United States after 1995 was primarily driven by the increased use of information capital goods. In addition, Vu (2011) finds that the impact of internet use on economic growth is far greater than that of hardware, such as mobile phones and personal computers.

To date, however, culture's role in this process has received little attention from researchers. We believe that culture is an essential factor to be considered in the social development process. For instance, the mitigation of COVID-19 seems to have been better facilitated in East Asian societies, which emphasize a collectivistic culture, rather than in Western countries, where culture is increasingly individualistic (Gokmen et al., 2021; Gupta et al., 2021). Weber (1930) claims that the development of Capitalism was fundamentally influenced by cultural transformations, particularly the Calvinist Reformation. Furthermore, Lal (2001) shows that the rapid growth of free markets in Western countries was determined by their individualistic culture. Landes (2000) argues that differences in economic development are shaped by culture. In this new era of social development, therefore, it is of great importance to examine how national culture is related to the development of the digital assets of a society.

In this chapter, we study whether national culture impacts the use of digital assets. To do so, we start with Hofstede et al.'s (2010) six cultural dimensions.[1] Next, we collect data from the World Bank and employ

[1] Hofstede et al. (2010) summarize the differences between cultures into six dimensions: (1) Power distance, (2) Uncertainty avoidance, (3) Masculinity, (4) Individualism, (5) Long-term orientation, and (6) Indulgence. This study has long been considered one of the classics in the field of cultural studies. Among them, (1) Power distance refers to the degree to which society accepts the unequal distribution of power. The higher the value of this dimension, the higher the acceptance of such inequality. For example, Europeans and Americans value individual competence over power, so the United States scores 40 on this dimension. However, Asian countries pay more attention to the constraint of power, China scores 80 on this dimension. (2) Uncertainty avoidance refers to the degree to which people of the same culture feel threatened by an ambiguous, unknown

four ways to capture the use of digital assets (digital payments, mobile payments, using the internet to pay bills, and using the internet for online shopping). Our dataset contains 109 countries or regions. For the empirical analyzes, we start by discussing how financial institutions leverage big data. To capture the effect of national financial development, we then use credit card ownership and debit card ownership as proxies. We also use a set of variables as control variables, including residential savings, the land area of a country, the number of servers, the ratio of the labor force to the total population, and gross domestic product (GDP) per capita. For the empirical analyses, we use Hofstede et al.'s (2010) cultural dimensions as the main measure of culture and the Global Leadership and Organizational Behavior Effectiveness (GLOBE) project's (House et al., 2004) cultural indicators and the World Value Survey as alternative measures of culture for the robustness tests.

This chapter is organized as follows. Section 2 reviews the literature on the impact of the internet and society's digitalization on economic growth. Section 3 describes the data and research methods and proposes different hypotheses that this chapter will test. Section 4 tests how financial development affects the use of digital assets and presents the results for effects of culture on digital assets, using the hypotheses presented in the previous section. Finally, Sect. 5 provides a conclusion to the chapter.

situation. The higher a society scores on this dimension, the less tolerant it is of extreme thinking and behavior. In the face of uncertainty and unfamiliar situations, we will try our best to avoid and control uncertainty. On the contrary, a culture of low uncertainty avoidance tolerates deviant behavior and opinions, allowing different philosophical and religious views to coexist. (3) Masculinity depends on whether a society is more marked by traits that represent men, or more marked by traits that represent women. For example, competition and confidence represent masculine traits, whereas friendliness and mutual assistance represent feminine traits. The greater the masculinity index, the greater the degree to which a society is driven by success whereas a low measure on this index implies that the core values of society are kindness and concern for others. (4) Individualism measures whether a society is inclined to look out for the good of the individual or the good of the group. Societies with high individualism tend to be less connected, and high scores on this dimension indicate that people care more about their own lives. On the other hand, collectivist societies score lower and focus on intra-group relations and solidarity. (5) Long-term orientation refers to people's willingness to meet all their needs in the future. Higher scores on this dimension indicate that the culture of the community is pragmatic and frugal, saving for the future. Scholars point out that the long-term orientation is one of the main reasons for the rapid economic development in East Asia at the end of the twentieth century. (6) Indulgence is the extent to which people try to control their desires and behaviors. The higher the value, the more society allows self-indulgence, and the fewer constraints and restrictions people place on themselves.

2 Literature Review

The expansion of broadband internet has been shown by researchers to significantly influence the development of the local economy. Kenny (2002) finds that information and communications technology (ICT) facilitates production and trade for income generation, and the effect is even more powerful in poor and developing countries. Freund and Weinhold (2004) investigate the effect of the internet on international trade during the period from 1997 to 1999, showing that the growth of web hosts has had a considerable influence on trade growth. Furthermore, Holt and Jamison (2009) analyze the relations between ICT and local economic growth as well as between broadband internet and economic growth. The authors show that both relations are positive. Using a panel of Organisation for Economic Co-operation and Development (OECD) countries from 1996 to 2007, Czernich et al. (2011) find that the increase in high-speed broadband penetration raises the per capita annual economic growth rate. Additionally, Kolko (2012) finds a positive relationship between internet usage and local economic and employment growth. He finds this effect more substantial in industries that rely more on information technology and regions with lower population densities. A further study by Salahuddin and Gow (2016) shows that the use of the internet and traditional financial development have led to South Africa's economic growth from 1991 to 2013. Overall, the above studies indicate that the popularization of the internet has facilitated growth in international trade and has contributed to local economic development. Both of these effects have thereby led to global economic growth.

The latest internet finance research hotspots used to measure digital assets are mobile payments and internet shopping (Liu et al., 2020). The degree to which a society accepts digital assets is of great importance in determining economic growth. Dahlberg et al. (2008) reviewed the literature on mobile payments by analyzing the various factors that influence mobile payment-service markets and showed that the consumer perspective of mobile payments and the technical security of mobile payments are the two areas best covered by contemporary research. However, they also find that the effects of social and cultural factors on mobile payments and comparisons between mobile and traditional payment services remain uninvestigated issues.

Subsequent studies have filled some of this research gap. For example, Liébana-Cabanillas et al. (2014) studied the influence of gender on the acceptance of mobile payments, finding that the decisive factors (external influences, ease of use, attitude, usefulness, trust, and risk) have different effects on the acceptance of a mobile payment system when the gender of the consumer is considered. Furthermore, De Reuver et al. (2015) argue that social factors are important in this discussion because mobile payments require collaboration between financial institutions and internet operators. They find that distinct strategic goals and interests, conflicts, lack of dependency, and governance issues have led to the disintegration of mobile payment platforms. To date, however, there seems to exist no study that explicitly tests the effect of national culture on the use of digital assets, leaving the question open to discussion, which we will address in our chapter.

3 Methodology

3.1 Hypothesis Development

Financial institutions are leveraging the internet and big data to induce consumers' greater usage of digital assets. Trelewicz (2017) extensively discusses the relevance of big data to financial markets and corporate banking, including corporate credit and trading. The author outlines future opportunities regarding how new information technologies entail massive volumes of transactions and money. In a recent study, Hasan et al. (2020) systematically review how big data connects to finance in the age of technology. The authors argue that big data significantly influences the financial field, as there is an abundance of transactions occurring daily that financial institutions are collecting as data. In turn, big data is shaping financial products and services. Moreover, Hasan et al. (2020) provide an extensive discussion on how big data influences financial markets, financial institutions, internet finance, financial management, internet credit service companies, fraud detection, risk analysis, and financial application management, which demonstrates the many ways in which big data is having an impact on finance.

In summary, the literature suggests that modern finance connects deeply to big data. Therefore, our first hypothesis is:

H1: Financial institution development is positively associated with using digital assets.

Additionally, culture is one of the most critical factors influencing human behavior (Ford, 1942; Hofstede, 1980, 2001). National culture can be viewed as the country's common values. Although scholars have done much research on how national cultural traits impact the development of financial institutions and economies, there is very little in the literature on the relationship between national culture and digital assets. From the economic development perspective, cultural characteristics contribute unevenly to economic growth. In other words, cultural values may facilitate or hinder economic growth (Castellani, 2019). Papamarcos and Watson (2006) develop an interactive model that shows that culture significantly influences differences in per capita GDP growth across countries.

Moreover, purposeful action, based on inner beliefs, is a driving force of economic growth. For example, Maridal (2013) cites that the speed of economic growth is higher in countries where people behave more honestly. Additionally, the cultural dimensions of Hofstede et al. (2010) have been shown in the literature to have significant relevance to economic development (Kristjánsdóttir et al., 2017; Kwok & Tadesse, 2006; Lavezzolo et al., 2018).

Despite these findings, however, a research gap exists in the relation between national culture and the use of digital assets, which we address in this chapter. We argue that the use of digital assets connects with Hofstede et al.'s (2010) culture dimensions as follows:

(1) **Uncertainty Avoidance Index (UAI)**: A lower score means people have strong adaptability and entrepreneurial spirit. This spirit may encourage more entrepreneurs to invest and innovate regarding digital assets.
(2) **Long Term Orientation Index (LTO)**: A higher score suggests an ability to adapt easily to a change in conditions, a strong propensity to save and invest, thriftiness, and perseverance in achieving results. As a result of new advancements in internet technology, for example, more digital assets are likely to be developed.
(3) **Individualism Index (IDV)**: A highly collectivist (lower individualistic) culture implies that people act in the interests of the group over their own self-interest. The wide application of digital assets may therefore encourage people to use them to integrate with other members of society.
(4) **Power Distance Index (PDI)**: A higher score indicates that society believes that inequalities amongst people are acceptable. People should, therefore, not seek to advance their self-interests over the

group's interests. Whether this factor impacts the use of digital assets is difficult to predict.

(5) **Indulgence Index (IVR)**: If a society that values restraint has a lower score in this dimension, people are bound by social norms and feel it is wrong to indulge themselves. If the indulgence score is high, people may be more inclined to increase consumption and take out loans through digital assets.

(6) **Masculinity Index (MAS)**: A masculine society values work over family and leisure activities. Under this condition, people are more inclined to use convenient and fast digital assets to engage in complicated economic activities, such as deposits, loans, and investments.

We summarize the above hypotheses as follows:

H2a: Uncertainty avoidance and the use of digital assets are negatively related.

H2b: Long-term orientation and the use of digital assets are positively related.

H2c: Individualism and the use of digital assets are negatively related.

H2d: Power distance and the use of digital assets are insignificantly related.

H2e: Indulgence and the use of digital assets are positively related.

H2f: Masculinity and the use of digital assets are positively related.

Trust is one national cultural trait emphasized in the literature but not captured by Hofstede et al.'s (2010) cultural dimensions. Zak and Knack (2001) document that high trust environments increase the rate of investment, promoting the economy's growth. Bottazzi et al. (2016) report that trust positively affects financial investment, particularly in contracting decisions involving contingent control rights. The increased use of digital assets generates a considerable amount of customer data, and so firms increasingly rely on the trustworthiness of the obtained data. As Culnan and Armstrong (1999) point out, the private information of individuals provides value to organizations, and so it follows that the wide use of big data depends largely on the social culture of trust. Formally, we define an additional hypothesis as follows:

H3: Trust and the use of digital assets are positively related.

3.2 Data, Variables, and Modeling

Data used in this study was obtained from the World Bank database. The initial panel includes 206 countries or regions around the world. However, not all these countries (regions) have cultural and/or digital assets data available. In the analysis of the impact of culture on digital assets, we focus on 109 countries or regions. As the critical variable of interest, researchers can capture national culture in several ways. In this chapter, we mainly use Hofstede et al.'s (2010) cultural dimensions, as proposed by Dutch psychologist Geert Hofstede, to study the correlation between culture and the development of the internet economy. Hofstede defines culture as a collective way of thinking that distinguishes the members of one group of people from another. Hofstede (1980, 2001) and Hofstede et al., (2010) have conducted an exhaustive study of national values and generated a dimensional paradigm and summarized the differences between cultures into six basic value dimensions: *UAI, LTO, IDV, PDI, IVR,* and *MAS*. The measurement of the degree of each dimension adopts a scoring system between 0 and 100 (low to high).

As a robustness check, we also consider the cultural values definitions developed by the GLOBE project (House et al., 2004). Furthermore, we consider the cultural trait of trust, which is captured by neither Hofstede's project nor the Globe's. Per the existing literature (Ahern et al., 2015; Guiso et al., 2006), we measure trust based on the World Value Survey (WVS) question, "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" We normalize the average answer to the questions to be bounded between zero and one for the countries in the sample.

For digital assets, we use four measures as proxies: the use of digital assets as measured by digital payments (*DigPay*), mobile payments (*MobPay*), using the internet to pay bills (*IntBill*), and using the internet for online shopping (*IntBuy*). The World Bank database contains data only from 2017. For the purpose of this study, this data provides us with a more natural context as the economy was very stable in 2017 as opposed to other years around this period.[2] Furthermore, the 2017 data also fits our cross-sectional analysis requirement because national culture does not change over time. In all ordinary least squares (OLS) regressions conducted, we control for financial institutional development (using *Credit card ownership* and *Debit card ownership* which generate similar results), resident savings (*Saving*), the land area of the country (*Land area*), the number of servers per 1 million people (*Number of servers*), the ratio of the labor force to the total population (*Labor Population*), and *GDP per capita*. We provide definitions for all of these variables in the Appendix.

Table 1 provides summary statistics for the above-defined variables. The digital payment's mean (median) value is 51.43% (48.25%) with a standard deviation of 29.88%. Notably, this is an aggregate measure that includes mobile payments, internet bill payments, and internet shopping. The number suggests that around half the population in our sample uses digital payments. On average, people who own a credit card represent less than a quarter of the total population in our sample, while around a half own a debit card. These figures indicate there is still room for growth in digital assets and financial development.

4 Results

4.1 Financial Institutions and the Use of Digital Assets

We begin our analysis by exploring how financial institutions leverage big data to encourage greater use of digital assets. Specifically, we estimate the following equation:

$$\text{DigAssets}_i = \alpha + \beta \cdot \text{FinInstitution}_i + \text{Controls}_i + \varepsilon \tag{1}$$

where DigAssets_i represents the use of digital assets in the country i measured by the overall digital payments (*DigPay*), mobile payments (*MobPay*), use of the internet to pay bills (*IntBill*), and online shopping (*IntBuy*). FinInstitution_i stands for the financial development of the

[2] This helps exclude the influence of abnormal economic conditions such as financial crises or economic bubbles. Like Nicholas Kristoff of the New York Times: "Why 2017 Was the Best Year in Human History" he wrote that a smaller share of the world's people was hungry, impoverished, or illiterate than at any time before. A smaller proportion of children died than ever before.

Table 1 Summary statistics

Variables	N	*Mean*	*Stdev.*	*p10*	*p50*	*p90*
DigPay	220	51.43	29.88	13.98	48.25	94.80
MobPay	220	7.57	7.86	0.87	4.72	18.33
IntBill	218	25.66	24.67	2.70	12.28	64.61
IntBuy	218	24.50	23.47	1.59	16.47	66.84
Credit card ownership	220	22.65	22.03	1.45	14.76	61.83
Debit card ownership	220	49.80	31.55	9.54	50.54	91.65
UAI	168	64.82	20.76	36.00	64.50	92.00
LTO	168	41.01	23.54	13.00	35.00	75.00
IDV	168	40.50	22.95	16.00	35.50	75.00
PDI	168	61.68	20.29	34.00	64.00	80.00
IVR	168	46.87	21.94	18.00	42.50	78.00
MAS	168	48.48	17.87	26.00	46.50	68.00
Trust	168	0.33	0.22	0.08	0.28	0.66
Saving	220	52.73	17.21	30.20	52.74	77.33
Land area	220	12.04	2.04	9.85	12.21	14.41
Number of servers	220	6.42	3.03	1.81	6.86	10.00
Labor Population	220	45.90	8.11	34.43	46.99	54.98
GDP per capita	216	8.93	1.47	6.72	8.96	10.76

Note This table reports the summary statistics for the variables used in this paper. *DigPay* represents digital payments, *MobPay* represents mobile payments, *IntBill* represents using the Internet to pay bills, *IntBuy* represents using the Internet for online shopping, *Credit card ownership* and *Debit card ownership* proxy for country-level financial institutional development, *UAI* represents Uncertainty Avoidance, *LTO* represents Long-Term Orientation, *IDV* represents Individualism, *PDI* represents Power Distance, *IVR* represents Indulgence, and *MAS* represents Masculinity. *Trust* is measured by asking the following WVS question: "*Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?*" and normalizing the average answer to be bounded between zero and one for all countries in the sample. *Saving* represents resident savings, *Land area* represents the natural logarithm of the land area of a country (km^2), *Number of servers* represents the number of servers per 1 million people, *Labor Population* represents the ratio of labor force to the total population, and *GDP per capita* represents per capita gross domestic product. All these variables are defined in the Appendix. All data points are obtained from the World Bank

country i proxied by credit card ownership and debit card ownership. The control variables include *Saving*, *Land area*, *Number of servers*, *Labor Population*, and *GDP per capita*. As mentioned earlier, we carry out this estimation using a cross-sectional analysis. To ensure the regression results are unbiased, we cluster standard errors on the country level and ensure robustness to heteroscedasticity.

We report the results with the OLS specifications in Table 2, where columns (1) to (4) use *Credit card ownership* and columns (5) to (8) use *Debit card ownership* to capture financial institutional development. The coefficients on both measures are positive and statistically significant throughout the columns at the 5% level or better. This result confirms that the development of financial institutions at the country level contributes to the use of digital assets.

As for control variables, the results are also interesting. First, individual savings are consistently and positively related to the use of digital assets. Secondly, land area is significantly and positively related to mobile payments and online shopping but not significantly associated with paying bills online. Third, the number of internet servers appears to be significant in the regressions where we use credit card ownership to capture financial institutions but insignificant in the regression where we use debit card ownership to capture financial institutional development. Fourth, the labor force ratio is insignificant when related to the use of digital assets (we use the total population to replace labor force and obtain the same results). Finally, GDP per capita is significantly related to the use of digital assets for online shopping but insignificant for other purposes.

In all regressions, the adjusted R^2 ranges between 0.42 and 0.87, indicating that these variables demonstrate a good explanatory power for the use of digital assets. In particular, using debit card ownership to measure financial development shows a better explanatory power than using credit card ownership. Overall, the results in this subsection provide evidence supporting hypothesis H1.

4.2 The Role of Culture in the Use of Digital Assets

4.2.1 Evidence from Hofstede Culture Dimensions

In this section, we examine our hypotheses on the relationship between national culture and the use of digital assets. Empirically, we estimate the following equation:

$$\text{DigAssets}_i = \alpha + \beta_1 \text{Culture}_i + \beta_2 \text{FinInstitution}_i + \gamma_c \text{Controls}_i + \varepsilon_i \quad (2)$$

where DigAssets_i represents the use of digital assets in the country i measured by the overall digital payments (*DigPay*), mobile payments (*MobPay*), usage of the internet to pay bills (*IntBill*), and online shopping (*IntBuy*), and $Culture_i$ represents Hofstede et al.'s (2010) cultural dimensions: *UAI, LTO, IDV, PDI, IVR*, and *MAS*. We also control for financial institutional development (*credit card ownership* and *debit card*

Table 2 Use of digital assets at financial institutions

Variables	*DigPay* (1)	*MobPay* (2)	*IntBill* (3)	*IntBuy* (4)	*DigPay* (5)	*MobPay* (6)	*IntBill* (7)	*IntBuy* (8)
Credit card ownership	0.479***	0.111**	0.356**	0.394***				
	(4.270)	(2.190)	(2.590)	(3.880)				
Debit card ownership					0.705***	0.160***	0.497***	0.342***
					(4.770)	(2.790)	(5.600)	(4.720)
Saving	0.475***	0.178***	0.421***	0.359***	0.466***	0.177***	0.419***	0.412***
	(4.990)	(3.590)	(5.110)	(5.510)	(6.160)	(4.090)	(6.380)	(7.590)
Land area	−0.254	1.022***	0.671	1.075**	−0.148	1.048***	0.734	1.204***
	(−0.440)	(4.040)	(1.240)	(2.540)	(−0.330)	(4.240)	(1.310)	(2.870)
Number of servers	2.571***	0.109	1.963**	1.546***	0.190	−0.431	0.283	0.338
	(2.850)	(0.320)	(2.370)	(2.640)	(0.240)	(−1.220)	(0.330)	(0.430)
Labor Population	0.192	0.048	0.054	0.096	0.024	0.010	−0.062	0.030
	(0.860)	(0.550)	(0.370)	(0.880)	(0.130)	(0.120)	(−0.470)	(0.280)
GDP per capita	2.336	−0.163	2.850	3.701***	−0.205	−0.711	1.251	4.261***
	(0.940)	(−0.160)	(1.600)	(3.100)	(−0.070)	(−0.630)	(0.650)	(2.820)
Constant	−27.809	−18.102**	−53.514***	−63.747***	−7.089	−13.673	−40.356**	−70.317***
	(−1.530)	(−2.360)	(−3.320)	(−4.870)	(−0.380)	(−1.540)	(−2.370)	(−5.550)
Adjusted R^2	0.812	0.427	0.767	0.869	0.870	0.469	0.806	0.870
Observations	109	109	108	108	109	109	108	108

Note This table reports the OLS regression results for the country-level financial institutional development, proxied by credit card ownership and debit card ownership, on the use of digital assets, measured by digital payments (*DigPay*), mobile payments (*MobPay*), using Internet to pay bills (*IntBill*), and using Internet for online shopping (*IntBuy*), while controlling for resident savings (*Saving*), the land area of a country (*Land area*), the number of servers per 1 million people (*Number of servers*), the ratio of labor force to the total population (*Labor Population*), and *GDP per capita*. All these variables are defined in the Appendix. All data points are obtained from the World Bank. The standard errors are clustered on the country level and are robust to heteroskedasticity. The values of t-statistics are reported in parentheses. ***, **, and * represent the 1%, 5%, and 10% level of significance, respectively *ownership*), *Saving*, *Land area*, *Number of servers*, *Labor population*, and *GDP per capita*. To ensure the absence of bias in the regression results, we use robust standard errors and cluster them at the country level.

We report the results for specification (2) in Table 3. Firstly, in terms of Uncertainty Avoidance (*UAI*), we find a negative effect in the mobile payments (*MobPay*) regression. In all other regressions, however, the results are insignificant. This result offers weak support for hypothesis H2a. In terms of Long-Term Orientation (*LTO*), we find a positive effect on overall digital payments (*DigPay*) and using the internet for online shopping (*IntBuy*). Nonetheless, we do not find significant results regarding mobile payments (*MobPay*) and using the internet to pay bills (*IntBill*). Thus, in general terms, we can only claim that the evidence partially supports hypothesis H2b.

The results obtained for Individualism (*IDV*) are interesting. In the hypothesis development section, we argue that individualism should be negatively related to the use of digital assets because a more highly individualistic culture implies that people act more from their own self-interest than that of the group. The wide application of digital assets may therefore not apply to individualistic cultures, which do not emphasize a greater degree of integration with other members of society. However,

Table 3 National culture and the use of digital assets

Variables	*UAI*	*LTO*	*IDV*	*PDI*	*IVR*	*MAS*
DigPay	×	+	+	×	×	×
MobPay	—	×	+	×	×	—
IntBill	×	×	+	×	—	—
IntBuy	×	+	+	×	×	×

Note This table summarizes the test results for the relationship between the use of digital assets and Hofstede's cultural dimensions, with + , —, and × representing a significantly positive result at least at the 10% level, a significantly negative result at least at the 10% level, and insignificance at the 10% level, respectively. The use of digital assets is measured by digital payments (*DigPay*), mobile payments (*MobPay*), using the Internet to pay bills (*IntBill*), and using the Internet for online shopping (*IntBuy*). For cultural dimensions, *UAI* stands for Uncertainty Avoidance, *LTO* stands for Long-Term Orientation, *IDV* stands for Individualism, *PDI* stands for Power Distance, *IVR* stands for Indulgence, and *MAS* stands for Masculinity. In all regressions, we control for financial institutional development (using *Credit card ownership* and *Debit card ownership* generate similar results), resident savings (*Saving*), the land area of a country (*Land area*), the number of servers per 1 million people (*Number of servers*), the ratio of labor force to the total population (*Labor Population*), and *GDP per capita*. All data points are obtained from the World Bank. All the variables are defined in the Appendix the evidence shows that *IDV* is positively, significantly, and persistently related to digital assets in all four regressions. A plausible explanation for this is that people in individualistic cultures are more innovative, and so they are not held back by a traditional way of doing things. Therefore, they are more likely to use digital assets when such are available. Based on the evidence from this analysis, we therefore reject hypothesis H2c. Additionally, we carry out robustness tests with alternative cultural measures and discuss these results in more detail in the next section.

In terms of the power distance (*PDI*), the results are consistent with our expectations. With all four measures of the use of digital assets, we do not find any significant evidence that they are associated with *PDI*. Hence, the null hypothesis of H2d cannot be rejected. For masculinity (*MAS*) and indulgence (*IND*), coefficients are negative for using the internet to pay bills (*IntBill*), and masculinity (*MAS*) is negatively associated with mobile payments (*MobPay*). According to these results, hypotheses H2e and H2f appear to be rejected. However, the results are not statistically significant for other ways of using digital assets. In particular, the evidence is not significant for overall digital payments. Therefore, the last two hypotheses are not supported, and the effect of masculinity and indulgence on the use of digital assets is not apparent.

Overall, the results in this subsection indicate that in light of Hofstede et al.'s (2010) cultural dimensions, only individualism (*IDV*) has a consistently positive effect on the use of digital assets. The results obtained regarding other cultural dimensions are generally mixed.

4.2.2 Evidence from Alternative Measures of Culture

The above subsection shows that individualism is positively related to the use of digital assets. However, Hofstede et al.'s (2010) cultural framework faces criticism in the literature (Baskerville, 2003). In this section, we use alternative cultural measures to test the results' robustness further. In particular, we use the GLOBE project's (House et al., 2004) societal collectivism indicator, which is similar to Hofstede et al.'s individualism dimension but captures the opposite effect. While Hofstede (2006) criticizes the GLOBE's cultural indicators because of how they are measured, these indicators are commonly used in the literature (Karolyi, 2016; So & Zhang, 2022; Zhang, 2022).

The results reported in Panel A of Table 4 show that the coefficients on collectivism are negative in seven out of eight regressions, in which six coefficients are significant at the 10% level or better. These results suggest that collectivism negatively relates to the use of digital assets and, in essence, confirms our results in the previous subsection that individualism positively relates to the use of digital assets. Therefore, while GLOBE's cultural indicators and Hofstede et al.'s (2010) cultural dimensions are constructed respectively using different methodologies, they produce consistent results for the use of digital assets.

One issue worthy of further investigation is how trust, along with individualism, influences the use of digital assets, as trust represents another important cultural trait (Doney et al., 1998; Zak & Knack, 2001). A higher level of individualism does not mean a lower level of trust. Similarly, a higher level of collectivism does not mean a higher level of trust. The tendency for individuals in a society to personalize their behaviors or thoughts does not mean they are unwilling to cooperate with others. In fact, Van Hoorn (2014) finds that the opposite is true. That is, a higher level of individualism is associated with a higher level of trust. We report the results of tests regarding the effect of trust in Panel B of Table 4. These results show that the coefficients of trust are positive in seven out of eight regressions, in which six coefficients are statistically significant at the 1% level. These results suggest that trust is positively related to the use of digital assets.

In short, the results in this subsection confirm that individualism positively relates to the use of digital assets, supporting the results in the previous section. Furthermore, social trust exerts a significantly positive effect on the use of digital assets, providing support for Hypothesis 3.

5 Conclusion

In this chapter, we examine the influence of culture on the use of digital assets. Based on the relevant data from the World Bank database, we first use a cross-sectional analysis to study the relationship between financial institutional development and the use of digital assets. We find that they are significantly and positively related. We then make connections between national cultures and the use of digital assets. To identify which cultural traits are most significantly associated with the use of digital assets, we examine the six dimensions of Hofstede et al.'s (2010) cultural variables. We find that the individualism dimension has the greatest influence on the use of digital assets. Additionally, we consider the GLOBE project's (2004) measure of culture as a robustness test, showing that

Table 4 Use of digital assets at financial institutions while controlling for cultural collectivism (Panel A) and trust (Panel B)

Variables	*DigPay* (1)	*MobPay* (2)	*IntBill* (3)	*IntBuy* (4)	*DigPay* (5)	*MobPay* (6)	*IntBill* (7)	*IntBuy* (8)
Panel A: GLOBE Collectivism Value								
Collectivism	2.767	−4.935**	−7.999**	−6.532*	−1.619	−5.572***	−10.656***	−8.755***
	(0.770)	(−2.440)	(−2.110)	(−1.830)	(−0.640)	(−2.890)	(−3.110)	(−2.810)
Credit card ownership	0.303**	0.058	0.099	0.093				
	(2.330)	(0.780)	(0.540)	(0.620)				
Debit card ownership					0.476***	0.040	0.461***	0.364***
					(2.800)	(0.750)	(4.580)	(3.670)
Saving	0.435***	0.115**	0.316***	0.376***	0.379***	0.115**	0.231***	0.312***
	(4.00)	(2.370)	(3.440)	(5.100)	(3.650)	(2.310)	(2.930)	(4.510)
Land area	−0.335	1.399***	0.919	1.351**	−0.113	1.466***	0.844*	1.314***
	(−0.430)	(4.260)	(1.400)	(2.150)	(−0.220)	(4.700)	(1.730)	(2.710)
Number of servers	3.544**	−0.084	0.949	0.297	0.960	−0.350	−1.265	−1.474
	(2.650)	(−0.120)	(0.540)	(0.150)	(0.910)	(−0.460)	(−0.830)	(−0.750)
Labor Population	0.375	0.059	0.026	0.020	0.045	0.026	−0.261	−0.210
	(1.200)	(0.520)	(0.100)	(0.090)	(0.150)	(0.200)	(−1.130)	(−1.210)
GDP per capita	5.962	1.815	11.681***	13.118***	7.411	2.576	9.267***	11.506***
	(1.200)	(0.830)	(3.310)	(3.570)	(1.540)	(1.590)	(3.390)	(4.080)
Constant	−84.786***	−13.284	−82.443***	−105.89***	−61.450**	−15.388	−35.452	−70.655***
	(−2.830)	(−0.910)	(−2.860)	(−3.720)	(−2.400)	(−1.120)	(−1.630)	(−3.640)

Variables	*DigPay* (1)	*MobPay* (2)	*IntBill* (3)	*IntBuy* (4)	*DigPay* (5)	*MobPay* (6)	*IntBill* (7)	*IntBuy* (8)
Adjusted R^2	0.894	0.623	0.811	0.867	0.920	0.620	0.857	0.895
Observations	59	59	59	59	59	59	59	59
Panel B: WVS Trust Value								
Trust	4.237	9.776***	25.576***	18.640***	−4.969	8.589***	19.581***	15.168***
	(0.520)	(2.900)	(3.730)	(4.140)	(−0.710)	(2.650)	(2.780)	(3.360)
Credit card ownership	0.423***	0.135***	0.253*	0.286***				
	(3.640)	(2.740)	(1.760)	(2.650)				
Debit card ownership					0.669***	0.094	0.419***	0.258***
					(4.000)	(1.360)	(4.010)	(3.100)
Saving	0.438***	0.125***	0.353***	0.355***	0.448***	0.148***	0.353***	0.393***
	(4.720)	(3.420)	(4.440)	(6.220)	(4.970)	(3.500)	(4.910)	(6.770)
Land area	−0.565	0.968***	0.051	0.456	−0.383	1.073***	0.123	0.643
	(−0.880)	(3.840)	(0.100)	(0.950)	(−0.720)	(3.580)	(0.240)	(1.450)
Number of servers	2.599**	−0.022	2.431**	1.641**	0.357	−0.324	1.021	0.796
	(2.47)	(−0.060)	(2.360)	(2.00)	(0.350)	(−0.700)	(0.920)	(0.720)
Labor Population	0.302	0.040	0.088	0.126	0.192	0.033	0.020	0.099
	(1.05)	(0.360)	(0.440)	(0.960)	(0.790)	(0.280)	(0.100)	(0.730)
GDP per capita	2.787	−0.661	2.518	4.252***	0.827	0.011	1.150	5.039***

(continued)

Table 4 (continued)

Variables	*DigPay* (1)	*MobPay* (2)	*IntBill* (3)	*IntBuy* (4)	*DigPay* (5)	*MobPay* (6)	*IntBill* (7)	*IntBuy* (8)
	(0.910)	(−0.590)	(1.310)	(3.360)	(0.240)	(0.010)	(0.530)	(3.210)
Constant	−32.290	−12.923	−50.712***	−66.940***	−18.034	−20.361*	−39.840**	−76.570***
	(−1.490)	(−1.640)	(−3.210)	(−5.240)	(−0.810)	(−1.830)	(−2.480)	(−6.510)
Adjusted R^2	0.806	0.554	0.809	0.903	0.861	0.541	0.838	0.903
Observations	92	92	91	91	92	92	91	91

Note This table reports the OLS regression results for the country-level financial institutional development of a country, proxied by *Credit card ownership* and *Debit card ownership*, on the use of digital assets, measured by digital payments (*DigPay*), mobile payments (*MobPay*), using the Internet to pay bills (*IntBill*), and using the Internet for online shopping (*IntBuy*), controlling for resident savings (*Saving*), the land area of a country (*Land area*), the number of servers per 1 million people (*Number of servers*), the ratio of labor force to the total population (*Labor Population*), and *GDP per capita*. All these variables are defined in the Appendix. All data points are obtained from the World Bank. The standard errors are clustered at the country level and are robust to heteroskedasticity. The values of t-statistics are reported in the parentheses. ***, **, and * represent the 1%, 5%, and 10% level of significance, respectively the value of collectivism significantly and negatively relates to the use of digital assets.

The results suggest that a cultural environment with a greater emphasis on personalization is conducive to the better development and greater use of digital assets. We further investigate the effect of social trust using the WVS data, which is also commonly cited in the literature (Guiso et al., 2006; Van Hoorn, 2014; Zak & Knack, 2001). We show that similar to individualism, trust positively affects digital assets' use. These findings provide evidence that certain cultural factors are indeed related to the use of digital assets.

We can draw some lessons from our analysis that will be of practical and academic use. Practitioners working in big data who want to understand where the industry may be heading need to first know how national culture is related to the use of digital assets. There are at least three reasons why this is so. Firstly, the acceleration of internationalization requires businesses in different countries to coordinate to be effective in their operations. This trend is intensified during COVID-19, as practitioners worked online to provide training or consultation in different countries. To serve the needs of diverse customers, it is imperative that data practitioners take measures to improve cultural competence (the ability to collaborate effectively with clients from different cultures), as such competence improves experiences and outcomes. Secondly, different cultures adopt new technologies differently and at different speeds. Only by knowing the traits and features rooted in different societies can a data practitioner help promote innovation in different cultures more validly and sustainably. Thirdly, it is also likely that a data practitioner is attracted to migrate to another country in the world of globalization. While the practitioner may contribute to the destination country's economy through his or her data skills, it is also essential for the overseas-qualified practitioners to self-adjust in terms of cultural awareness and understanding of the new working environment. In this chapter, our findings suggest that more attention should be paid to the levels of trust and individualism in society to determine if opportunities exist to more extensively promote the use of big data in business.

This chapter contributes to the literature on national culture, finance, and digital technology. As previously mentioned, due to the influence of COVID-19, closer attention is being paid to the development of the Internet industry and related digital industries. If a country's cultural values hinder the development of the Internet and related technologies, then the country's economic development may also be significantly affected. We, therefore, identify culture as an important driver of technological change and economic development. In the twenty-first century, the Internet has entered a stage of rapid development and has been integrated by society into our daily lives. Meanwhile, culture is a long-standing factor of social development for which there remain unanswered questions. For example, how can different cultures affect the development of digital assets and, in turn, facilitate the efficiency and equality of economies? Future research is needed to assess culture's influence on the digital economy's growth in the era of big data.

Acknowledgments We deeply appreciate feedback from Shaosong Wang and Shiyong Zhao, and valuable comments and suggestions from the editing team of Big Data in Finance—Opportunities and Challenges of Financial digitalization.

Appendix

Description of variables

Variable	*Description*	*Source*
Use of Digital Assets		
DigPay	The percentage of respondents who used electronic payments (payments that one makes or that are made automatically including wire transfers or payments made online) in the past 12 months to make payments on bills or to buy things using money from their accounts (% age 15+)	World Bank
MobPay	The percentage of respondents who report using a mobile phone to pay bills in the past 12 months (% age 15+)	World Bank
IntBill	The percentage of respondents who report using the Internet to pay bills in the past 12 months	World Bank

(continued)

(continued)

Variable	*Description*	*Source*
IntBuy	The percentage of respondents who report using the Internet to buy something online in the past year (% age 15+)	World Bank
Financial Institution Development		
Credit card ownership	The percentage of respondents with a credit card (% age 15+)	World Bank
Debit card ownership	The percentage of respondents with a debit card (% age 15+)	World Bank
National Cultural Traits		
Uncertainty avoidance, Long term orientation, Individualism, Power Distance, Indulgence, and Masculinity	Cultural dimensions as described in Hofstede et al. (2010)	The Hofstede Centre (https://geert-hofstede.com), and Geert Hofstede's academic website (http://www.geerthofstede.nl)
Collectivism	The societal collective indicator in House et al. (2004)	GLOBE project
Trust	The percentage to the answer of the World Value Survey (WVS) question, "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" The average answer is bounded between zero and one for all countries in the sample	World Value Survey (WVS)
Control Variables		
Saving	The percentage of respondents who report saving or setting aside any money in the past 12 months (% age 15+)	World Bank

(continued)

(continued)

Variable	*Description*	*Source*
Land area	Land area is the total area of the country, excluding area under inland water bodies, national claims to a continental shelf, and exclusive economic zones. In most cases the definition of inland water bodies includes major rivers and lakes	World Bank
Number of servers	The number of distinct, publicly trusted TLS/SSL certificates found in the Netcraft Secure Server Survey (per 1 million people)	World Bank
Labor Population	Labor force comprises people ages 15 and older who supply labor for the production of goods and services during a specified period. It includes people who are currently employed and people who are unemployed but seeking work as well as first-time jobseekers. Not everyone who works is included, however. Unpaid workers, family workers, and students are often omitted, and some countries do not count members of the armed forces. Labor force size tends to vary during the year as seasonal workers enter and leave	World Bank

(continued)

(continued)

Variable	*Description*	*Source*
GDP per capita	GDP per capita is gross domestic product divided by midyear population. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars	World Bank

References

Ahern, K. R., Daminelli, D., & Fracassi, C. (2015). Lost in translation? The effect of cultural values on mergers around the world. *Journal of Financial Economics, 117*(1), 165–189.

Baskerville, R. F. (2003). Hofstede never studied culture. *Accounting, Organizations and Society, 28*(1), 1–14.

Bottazzi, L., DaRin, M., & Hellmann, T. (2016). The importance of trust for investment: Evidence from venture capital. *Review of Financial Studies, 29*(9), 2283–2318.

Castellani, M. (2019). Does culture matter for the economic performance of countries? An overview of the literature. *Journal of Policy Modeling, 41*(4), 700–717.

Culnan, M. J., & Armstrong, P. K. (1999). Information privacy concerns, procedural fairness, and impersonal trust: An empirical investigation. *Organization Science, 10*(1), 104–115.

Czernich, N., Falck, O., Kretschmer, T., & Woessmann, L. (2011). Broadband infrastructure and economic growth. *The Economic Journal, 121*(552), 505–532.

Dahlberg, T., Mallat, N., Ondrus, J., & Zmijewska, A. (2008). Past, present and future of mobile payments research: A literature review. *Electronic Commerce Research and Applications, 7*(2), 165–181.

De Reuver, M., Verschuur, E., Nikayin, F., Cerpa, N., & Bouwman, H. (2015). Collective action for mobile payment platforms: A case study on collaboration issues between banks and telecom operators. *Electronic Commerce Research and Applications, 14*(5), 331–344.

Doney, P. M., Cannon, J. P., & Mullen, M. R. (1998). Understanding the influence of national culture on the development of trust. *Academy of Management Review, 23*(3), 601–620.

Ford, C. S. (1942). Culture and human behavior. *The Scientific Monthly, 55*(6), 546–557.

Freund, C. L., & Weinhold, D. (2004). The effect of the internet on international trade. *Journal of International Economics, 62*(1), 171–189.

Gokmen, Y., Baskici, C., & Ercil, Y. (2021). The impact of national culture on the increase of COVID-19: A cross-country analysis of European countries. *International Journal of Intercultural Relations, 81*, 1–8.

Guiso, L., Sapienza, P., & Zingales, L. (2006). Does culture affect economic outcomes? *Journal of Economic Perspectives, 20*(2), 23–48.

Gupta, M., Shoja, A., & Mikalef, P. (2021). Toward the understanding of national culture in the success of non-pharmaceutical technological interventions in mitigating COVID-19 pandemic. *Annals of Operations Research*.

Hasan, M. M., Popp, J., & Oláh, J. (2020). Current landscape and influence of big data on finance. *Journal of Big Data, 7*(1), 21.

Hofstede, G. H. (1980). *Culture's consequences: International differences in work-related values*. Sage.

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Sage.

Hofstede, G. H. (2006). What did GLOBE really measure? Researchers' minds versus respondents' minds. *Journal of International Business Studies, 37*(6), 882–896.

Hofstede, G. H. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online Readings in Psychology and Culture, 2*(1), 8.

Hofstede, G. H., Hofstede, G. J., & Minkov, M. (2010). *Cultures and organizations: Software of the mind: Intercultural cooperation and its importance for survival*. McGraw-Hill.

Holt, L., & Jamison, M. (2009). Broadband and contributions to economic growth: Lessons from the US experience. *Telecommunications Policy, 33*(10–11), 575–581.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., & Gupta, V. (2004). *Culture, leadership, and organizations: The GLOBE study of 62 societies*. Sage.

Jorgenson, D. W., Ho, M. S., & Stiroh, K. J. (2008). A retrospective look at the US productivity growth resurgence. *Journal of Economic Perspectives, 22*(1), 3–24.

Karolyi, G. A. (2016). The gravity of culture for finance. *Journal of Corporate Finance, 41*, 610–625.
Kenny, C. (2002). Information and communication technologies for direct poverty alleviation: Costs and benefits. *Development Policy Review, 20*(2), 141–157.
Kolko, J. (2012). Broadband and local growth. *Journal of Urban Economics, 71*(1), 100–113.
Kristjánsdóttir, H., Guðlaugsson, Þ. Ö., Guðmundsdóttir, S., & Aðalsteinsson, G. D. (2017). Hofstede national culture and international trade. *Applied Economics, 49*(57), 5792–5801.
Kwok, C. C. Y., & Tadesse, S. (2006). National culture and financial systems. *Journal of International Business Studies, 37*(2), 227–247.
Lal, D. (2001). *Unintended consequences: The impact of factor endowments, culture, and politics on long-run economic performance* (Vol. 7). MIT Press.
Landes, D. (2000). Culture makes almost all the difference. *Culture Matters: How Values Shape Human Progress*, 2–13.
Lavezzolo, S., Rodríguez-Lluesma, C., & Elvira, M. M. (2018). National culture and financial systems: The conditioning role of political context. *Journal of Business Research, 85*, 60–72.
Liébana-Cabanillas, F. J., Sánchez-Fernández, J., & Muñoz-Leiva, F. (2014). Role of gender on acceptance of mobile payment. *Industrial Management & Data Systems*.
Liu, J., Li, X., & Wang, S. (2020). What have we learnt from 10 years of fintech research? A scientometric analysis. *Technological Forecasting and Social Change, 155*, 120022.
Maridal, J. H. (2013). Cultural impact on national economic growth. *The Journal of Socio-Economics, 47*, 136–146.
Oliner, S. D., & Sichel, D. E. (2000). The resurgence of growth in the late 1990s: Is information technology the story? *Journal of Economic Perspectives, 14*(4), 3–22.
Osei, M. J., & Kim, J. (2020). Foreign direct investment and economic growth: Is more financial development better? *Economic Modelling, 93*, 154–161.
Papamarcos, S. D., & Watson, G. W. (2006). Culture's consequences for economic development: An empirical examination of culture, freedom, and national market performance. *Journal of Global Business and Technology, 2*(1), 48–57.
Salahuddin, M., & Gow, J. (2016). The effects of Internet usage, financial development and trade openness on economic growth in South Africa: A time series analysis. *Telematics and Informatics, 33*(4), 1141–1154.
So, J. Y., & Zhang, J. F. (2022). The effect of cultural heterogeneity on cash holdings of multinational businesses. *Research in International Business and Finance*, *61*, 101660.

Trelewicz, J. Q. (2017). Big data and big money: The role of data in the financial sector. *IT Professional, 19*(3), 8–10.
Van Hoorn, A. (2014). Individualist-collectivist culture and trust radius: A multilevel approach. *Journal of Cross-Cultural Psychology, 46*(2), 269–276.
Vu, K. M. (2011). ICT as a source of economic growth in the information age: Empirical evidence from the 1996–2005 period. *Telecommunications Policy, 35*(4), 357–372.
Weber, M. (1930). *The protestant ethic and the spirit of capitalism*. HarperCollins.
Xue, W. J. (2020). Financial sector development and growth volatility: An international study. *International Review of Economics & Finance, 70*, 67–88.
Zak, P. J., & Knack, S. (2001). Trust and growth. *The Economic Journal, 111*(470), 295–321.
Zhang, J. F. (2022). Cultural diversity and capital structures of multinational firms. *Review of Corporate Finance, 2*(2), 295–351.

Case Studies and Applications

Islamic Finance in Canada Powered by Big Data: A Case Study

Imran Abdool and Mustafa Abdool

1 Introduction

Credit unions are an important part of the Canadian financial ecosystem. Like banks, credit unions are deposit-taking institutions, but their ownership structure is different. In banks, depositors and shareholders are separate entities, i.e., when you make a deposit at a bank, you are not automatically an owner. To become an owner requires you to buy the bank's shares on the open market. In contrast, anyone using a credit union's services must first become a member by purchasing its membership shares. Credit unions are governed under a cooperative or mutual principle of "one member, one vote" (Ontario Co-operative Association, 2018). Therefore, all members have the same vote regardless of the deposit amount. While it is possible that members can purchase preferred shares, these shares do not confer additional voting power. However,

I. Abdool (✉)
Blue Krystal Technologies and Business Insights, Windsor, ON, Canada
e-mail: abdooli@mcmaster.ca

M. Abdool
Stanford University, Stanford, CA, USA
e-mail: moose878@stanford.edu

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_10

preferred shares could include preferential dividends or other privileges when compared to the standard membership shares. Membership in a credit union is restricted to those who fall within a “bond of association” (Credit Unions and Caisses Populaires Act, 1994). A bond of association is a common linkage among members such as ethnicity, profession, or geographic area. Ontario legislation requires that a bond of association link members. However, the legislation does permit a small percentage of members to be outside the bond of association at the discretion of the credit union’s Board of Directors.

Since 2010, a long-standing goal of the Greater Toronto Area’s (GTA) Muslim community has been the establishment of its own credit union. It is estimated that there are currently over 450,000 people self-identifying as Muslims in the GTA from the 2016 Statistics Canada census. To this end, the GTA Muslim community launched an initiative called the ICUC (Islamic Credit Union for Community). Motivations for starting a credit union can vary; sometimes prospective members believe that a credit union would increase cohesion within their community or further goals pertinent to the common bond of association. In the case of the proposed ICUC, the community’s motivation is to incorporate the principles of Islamic Finance into everyday banking services for Ontario’s Muslims.

The most important difference between Islamic Finance and the conventional system is its prohibition on interest. Islam, and some earlier iterations of the Judeo-Christian faith, claim “money should not beget money”, i.e., in order for money to earn a permissible return, it must be used in real economic activity. The second main difference between Islamic Finance and the conventional financial system is the prohibition of speculative transactions. For example, debt contracts cannot be traded in secondary markets. Therefore, an Islamic financial institution could not participate as a trader in the futures market. In Islamic Finance, the price for a transaction must always be known in advance and agreed to by the counterparty (Chapra, 2011).

Islamic financing, either in standalone institutions or a division within existing financial institutions, is quickly becoming common throughout the western world. For example, Hong Kong, Singapore, the United Kingdom, and Australia all now have significant Islamic financing hubs. There are two reasons for this. The first is a growing demand from Muslim communities for banking services compliant with their faith, and the second is that Islamic finance has been shown to have desirable characteristics for financial stability, at both the institution and system-wide level. For example:

- Two primary causes of the subprime mortgage crisis of 2007 were the improper verification of income by homebuyers and the contagion effect of misunderstood mortgage-backed securities. Islamic finance insists upon diligence by transacting parties. More so, mortgage-backed securities, being highly speculative by nature, could not exist on the balance sheet of Islamic financial institutions.
- The history of finance has long-standing examples of asset bubbles in financial markets bursting and causing significant damage to the real economy (output, employment, etc.). Such events include the dot.com bubble, the '90s Japanese real estate bubble, the Tulip bubble, and the South Seas bubble, among many others. While these occurred in different time periods and economies, leverage (or, as it is called today, margin trading) remains a common underlying factor. Islamic financial institutions cannot use margin trading as they fall under the earlier prohibition on speculation/gambling.

The ICUC has the support of all major Islamic centers in the GTA, such as the Toronto Islamic Center, the Islamic Information and Dawah Center, and the Islamic Society of North America, to name but a few. More granularly, several thousand prospective members have already completed an extensive membership survey. Currently, the community is engaged in high-level discussions with the Ontario financial regulator, the FSRA, and Ontario's Ministry of Finance. In November 2021, the community formally submitted its application.

Given the highly regulated sector in which credit unions operate, acquiring approval is a complex and multi-year endeavor. Determining the feasibility of a new credit union requires applying tools from econometrics and corporate finance, along with leveraging "big data" algorithms to make optimal decisions. This chapter aims to provide a case study on how big data and corporate finance are used to determine the feasibility of a new financial institution in Ontario.

2 Methods

As previously stated, starting a credit union requires a detailed feasibility study, business case, and business plan. The founders of a credit union must thoroughly convince the regulators on several matters. First, they must show that sufficient demand exists for their proposed credit union. Second, they must demonstrate that the credit union will meet the regulatory standards pertaining to risk and liquidity, more broadly known as capital adequacy. The current standard was set by the Basel Committee on Banking Supervision (an international committee responsible for coordinating and setting risk management for financial institutions) and is called "Basel III (Bank for International Settlements, 2017)". Third, if the credit union were to be unsuccessful, a plan to wind it down and minimize its losses must exist. This last point is especially important because deposit-taking financial institutions in Ontario (and Canada, more generally) are back-stopped by deposit insurance. This insurance guarantees depositors that their accounts are protected up to a predetermined limit (currently $250,000 for an Ontario licensed credit union).

As a practical matter, communities undertaking the creation of a new credit union start with a membership survey. The information collected from such a survey is critical as it provides the building blocks for the financial models. In turn, those models allow for scenario testing, forecasting, and other high-level analysis.

Given the foundational role of the membership survey, much thought must be given to its construction. Survey designers must balance acquiring sufficient information for robust financial models against the time and competing interests of survey-takers. A best practice is to explain to survey-takers that credit unions operate in a highly regulated sector. As such, this survey will be longer than what most survey-takers may be used to. In the case of the ICUC, completing its membership survey can take more than ten minutes. At a minimum, the membership survey should gather information pertaining to:

- Monthly cash in-flows and out-flows of members,
- Desired products and services,
- Capital structure, i.e., demand for preferred shares and at what price,
- Optimal location of branches,
- Community goals and their priority rankings,
- Interest by its members in volunteerism at the credit union,

- Governance rules within the bylaws that may require special attention, i.e., the process for revocation of membership, discounted share price for seniors, youth, etc.

Once the survey's raw data is collected, it must be cleaned, sorted, and inspected for its integrity. The modeling team should ensure that survey answers seem reasonable; for example, any statistical outliers should be identified and investigated. This step may require direct follow-up with the respondent to verify that their answer is indeed correct. Alternatively, in the interest of expediency, the modeling team may decide to automatically exclude answers beyond a certain threshold as outliers.

In the case of the ICUC's survey, one question inquired about a member's desired mortgage amount. Suppose the first two statistical moments (mean and median) for this question's answers are $650,000 and $600,000 respectively. Furthermore, suppose that one respondent's data point indicates a desired residential mortgage amount of $10,000,000—this would certainly be an outlier and require investigation. Was this answer just an error by the survey-taker, or does this prospective member intend to borrow such an amount? If the latter is true, it has important implications for the concentration of the credit union's portfolio of mortgages and the robustness of its risk management procedures.

Once the survey data has been collected, construction can begin on high-caliber dynamic and stochastic financial models. These models are among the best evidence that the proposed credit union can meet the regulatory standards and pass various stress-testing scenarios. In the absence of such models, decision-making would become a guessing game, and the success or failure of the financial institution would be left to chance. Even standard pro-forma financial statements, which by design are static, are insufficient for gaging the credit union's various risks and profit potential. Ultimately, the business environment is dynamic, and understanding how a prospective credit union would perform in a changing operating environment is indispensable. It should also be noted that these models are not just for regulatory use, as they also provide invaluable guidance to management. At a minimum, the models should be able to answer questions such as:

- What will be the impact of changes to the membership growth rate?
- What is the potential growth trajectory of the credit union's portfolio of assets?

- What is the Value at Risk (VAR) or Cumulative Value at Risk (CVAR)?
- What happens if there is a change in the amount of Tier 1 (safe assets) required?
- What is the maximum loanable amount that the credit union can lend out annually?
- What amount of non-performing loans can the credit union bear in an adverse scenario?
- What is the impact on return on equity (ROE), return on assets (ROA), and return on investment (ROI) as the credit union's portfolio of assets changes?
- What is a viable retention ratio, or alternatively, what might be the dividends the credit union can pay to members?

While detailed construction of financial models is beyond the scope of this chapter, a typical starting point would be to create a static pro-forma "3-statement" (i.e., income statement, balance sheet, and cash-flow statement) model. A 3-statement model links the income statement, balance sheet, and cash flow statement in one intricate financial model. Next, the model can become "dynamic", or time-varying as future values of line items become a function of time, i.e., the cash balance at time $t + n$ is then differentiated from that at time t. Finally, the model can be made stochastic: line-item future values are a function of some predictable component but also a random or shock component.

Choosing the exact specification for the model will depend on the context, historical data available, and best judgments about the future parameters and business operating environment. Of course, as new information becomes available, the model should be recalibrated to ensure its relevance and accuracy. Following the model's creation, simulations can be run to determine relevant ranges of output values, for example, what might be the cash balance of the credit union in ten years? These simulations are important because one aspect of the Basel III standards is that Tier I (safe assets) must make up a specific percentage of total assets. Running credible simulations instills confidence that the credit union will meet or exceed its regulatory thresholds. However, the building of state-of-the-art models requires the use of more sophisticated big data algorithms. Colloquially, if data is the new gold, then big data is the set of tools and techniques for extracting it.

Over the past decade, there has been an explosion of research and industry applications for big data and machine learning algorithms. From recommender systems in popular sites such as YouTube (Covington et al., 2016) to speech recognition models (Ning et al., 2019) installed on billions of mobile devices, it is not an understatement to say that big data has revolutionized key areas of our everyday life.

As one might expect, finance and the banking industry have also benefited greatly from state-of-the-art machine learning models. In the following sections, we give an overview of how an important sub-field of machine learning - deep learning—works. We offer a deep dive into how it has transformed two critical areas of the finance/banking industry: credit rating prediction and early detection of financial "stress" events.

3 Deep Learning Models

3.1 The Building Blocks of Deep Learning

The most general building block in deep learning is known as a node or neuron. A node takes as input several incoming signals, each with a different connection weight, and computes some output via a weighted summation. The final step is passing this output through a non-linear activation function (such as a sigmoid or tanh function). This non-linear transformation is crucial as it is what gives neural network models their expressive power and is also loosely inspired by the general workings of biological neurons in the human brain. Figure 1 provides an example of the mathematical representation of a single neuron:

We can create more sophisticated neural networks by stacking layers of nodes in sequence, which leads to structures as shown in Fig. 2 and what we refer to as a deep neural network (DNN). The process of training a deep learning model from data involves learning the connection weights between the nodes in each layer. During the training process, the connection weights are adjusted based on a specified loss function using a technique known as backpropagation (Rumelhart et al., 1986).

The idea of a loss function is critical to training DNNs. In a nutshell, this function measures how close the predicted value is to a ground truth label. For regression problems (e.g., where the predicted output is a real-valued number), the mean-squared-error loss function is a common

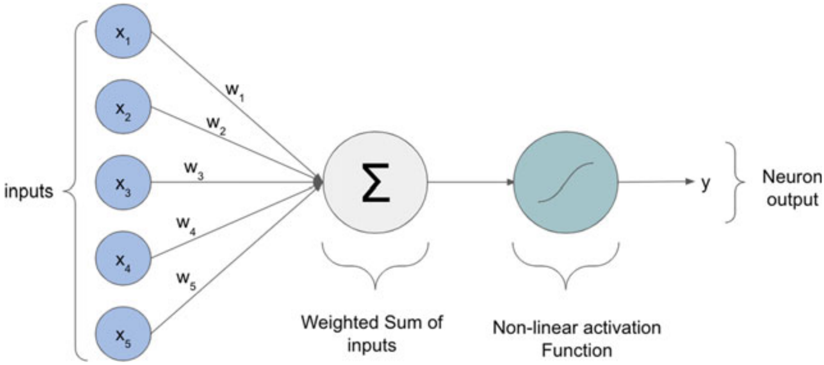


Fig. 1 A mathematical depiction of a single neuron as used in deep neural networks (*Note* A weighted-sum of inputs is computed and fed through a non-linear activation function to produce an output)

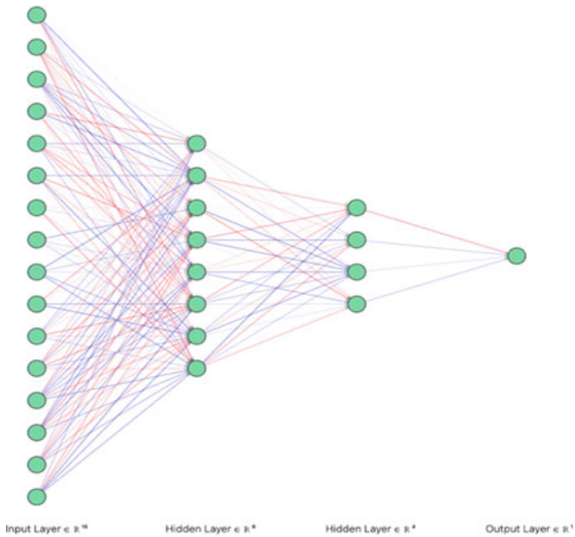


Fig. 2 Deep neural networks stacking architecture (*Note* In this figure, the green circles represent nodes, and the red and blue lines are the connection weights between nodes in different layers. In this network, sixteen inputs are transformed by hidden layers of size eight and four, respectively, to produce a single output)

choice. Mathematically, this loss function can be specified as shown in (1):

$$\mathrm{MSE}_{\mathrm{loss}} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \tag{1}$$

Equation 1: The Mean Squared Error Loss function as used in regression problems.

In this formulation, n is the number of total samples, y_i is the ground-truth label of the example and $\hat{y}_i$ is the predicted value from the model (a DNN in this case). The loss decreases as the predicted value becomes closer to the ground truth value.

Another widely used loss function is known as the cross-entropy loss. This type of loss function is used for classification problems, i.e., when the input is from a discrete set of possibilities. In the binary case, where we only have a positive and negative class, this loss function can be specified as shown in (2):

$$\mathrm{BCE}_{\mathrm{loss}} = -\frac{1}{n}\left(\sum_{i=1}^{n} y_i \log\left(\hat{y}_i\right) + (1 - y_i)\log\left(1 - \hat{y}_i\right)\right) \tag{2}$$

Equation 2: The definition of n, y_i, and $\hat{y}_i$ are the same as in the MSE loss function case. However, since we are dealing with a classification problem, we have the restriction that y_i can only be $\{0, 1\}$ which implies $\hat{y}_i$ represents the model's *predicted probability of the positive class*. Again, this loss function decreases as the predicted probability gets closer to the ground truth label.

For a simplified example of a deep learning model that a credit union might use, imagine trying to predict whether or not to approve an individual for a loan based on some relevant features (total household income, total amount of credit card debt etc.). In this case, the output of the model is simply two choices (approve loan/deny loan), so the binary cross-entropy loss function shown in Eq. 2 would be an appropriate choice.

To train such a model, we would need access to historical data with ground-truth labels. For example, we might have historical examples of real loans that were given, along with whether an individual was able to pay back the loans within the given terms or ended up defaulting. We could treat the repayment of the loan as the positive class ($y = 1$) and defaulting on the loan as the negative class ($y = 0$). Collecting all such examples would become the training data for our DNN model.

In practice, it is common to split the training data into a training set, validation set, and test set. Generally, we want most of the data to be used for training. Therefore, a 70-20-10 split for training, validation, and test sets respectively is commonly used. The data in the training set is used to learn the weights of the model based on a given loss function and the backpropagation algorithm, as previously mentioned. The examples in the validation set are never used to train, but rather are used to evaluate the model's ability to generalize to unseen examples during the training process. This set can help identify if the model is overfitting which, broadly speaking, means the model is simply memorizing patterns in the training data instead of generalizing. Overfitting is an important concept in machine learning, and a rich literature exists on techniques to mitigate it, such as reducing the model size or adding some type of regularization (Santos & Papa, 2022).

Lastly, the test set provides us with an unbiased way to evaluate the model once the training process is complete. Unlike the validation set, the test set is not even used to measure performance during training. Hence, the test set predictions are treated as the gold standard metric for a model's ability to generalize to new examples.

Nonetheless, this is just a surface-level view of the mechanics behind deep learning models and how to train them in practice; we strongly encourage readers to further explore the references to solidify their understanding. In the following sections, we explore concrete use-cases of how deep learning can be leveraged in credit unions, along with diving deeper into more sophisticated model architectures.

3.2 Deep Learning Models for Credit Scoring and Risk Prediction

One of the core functions of any banking institution is lending credit to consumers. The demand for consumer credit has generally been growing steadily from year to year. For example, in the United States, the federal reserve estimates that the current level of outstanding consumer credit amounts to approximately $4000 billion USD as of November 2021, which is increasing at a seasonally adjusted annual rate of 11% (US Federal Reverse, 2021). Given this demand, an active area of research is credit risk prediction, which queries whether or not to approve an individual for a personal loan or mortgage given their specific situation. As previously mentioned, this scenario of whether the institution approves or denies a loan based on particular features can be viewed as a binary classification problem. Naturally, the features should be at least somewhat relevant to the prediction task. An example of useful features might be: household income, total balance of all credit cards, number of credit cards, current balance of mortgage (if any), and current balance of car loans (if any). However, in practice, there could be hundreds of features, some of which may be highly correlated. As a general rule, keeping the set of features to a minimum is preferable as the addition of extraneous features affects the model's ability to generalize to unseen data (the overfitting problem mentioned in the previous section) and increases training time unnecessarily.

A recent paper that applies deep learning models for consumer credit scoring (Zhu et al., 2018) experiments using the popular Relief algorithm (Kira & Rendell, 1992) for feature selection. This algorithm automatically adjusts feature weights by observing what features are useful in discriminating between positive and negative outcomes. At the end of the algorithm, features that have a weight less than a specific threshold are pruned and removed before the training process begins. Their final set of experiments shows that deep learning models outperform traditional models (random forest and logistic regression) by a wide margin when used on a real-world dataset from a Chinese credit lending company.

3.3 Deep Learning Models for Processing Sequential Data

One major limitation of the above formulation is how it defines input features. Specifically, if one simply uses a fully-connected architecture, then all the sequential and chronological information is lost. To motivate the need for more sophisticated architectures, consider two example input features from the previous section: number of credit cards and balance of existing credit cards. If an applicant has three credit cards but just received two of them in the past month, then intuitively, they might be at a higher risk of defaulting on a personal loan. However, this important temporal information is lost if we only provide the model with two values that solely represent the number of credit cards and the total balance for all cards, respectively.

This idea of developing a deep learning architecture which can process sequential information, or a stream of events, has been previously studied. One of the most promising solutions is known as Recurrent Neural Networks or RNNs (Hochreiter & Schmidhuber, 1997). The main building block in such a network is a recurrent cell, a type of neuron with some form of memory that can be modified in response to an input. The overall idea is that a sequence of inputs can be fed, one at a time, into a sequence of recurrent cells which then performs some computation to update its weights. The key point is that each cell takes both the output of the previous cell and the next element in the sequence and uses both pieces of information to update its internal parameters. In this sense, the cell is able to "remember" the relevant data in the previous portion of the input sequence to solve the given task. Furthermore, similar to how neurons can be stacked in a basic neural network, we can stack layers of RNN cells to create even more expressive models. Figure 3 provides an example of a RNN architecture.

RNNs, specifically those using Long-Short-Term-Memory (LSTMs) cells, have achieved much success in recent years, especially in domains where the input and output data format is sequential. Some highlights include machine translation (Singh et al., 2017) and speech synthesis (Ning et al., 2019).

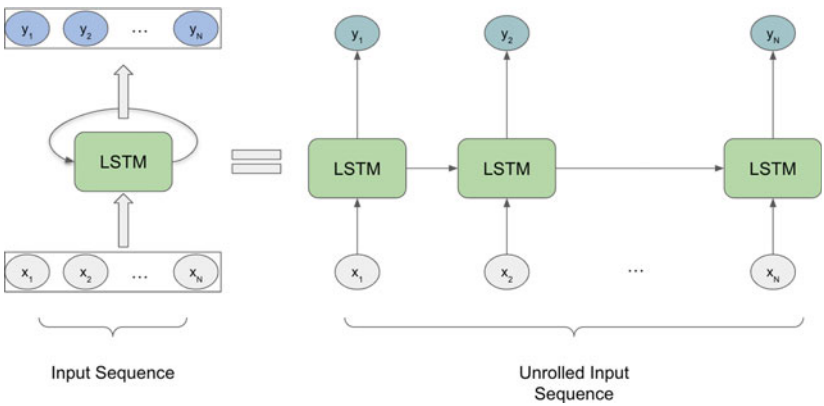


Fig. 3 Recurrent neural network (RNN) architecture using long short-term memory neural networks (LSTMs) (*Note* Recurrent Neural Network within an LSTM cell "unrolling" to process a sequence of inputs $[x_i]$ and produce a sequence of outputs $[y_i]$. Each subsequent LSTM cell incorporates both the current input and the output of the previous cell. Hence, in theory, the output of the last timestep can encode all the important information contained within the input sequence)

A further level of sophistication is augmenting the RNN with an attention mechanism (Bahdanau et al., 2014). The attention mechanism first gained popularity in neural machine translation to solve the problem that, in practice, RNNs tend to focus more on the contributions of later events due to the stronger gradient signals when using backpropagation. This bias toward later events can lead to subpar performance, especially in cases in which an important event occurred "early on" in the time-series (for example, if the first event an individual did was take out a credit card with a large limit).

The attention mechanism helps to mitigate this issue by allowing each RNN cell to learn custom weights over the previous input sequence. Essentially, this gives the model the flexibility to "focus" on events that are more relevant to the final prediction, independent of where they occur in the sequence. Similar to other advances in deep learning, the idea behind attention has its roots in human perception, specifically the notion that humans have been found to focus or "attend" to specific regions of an image when making decisions.

The idea behind the attention mechanism was taken even further with the development of transformer models (Vaswani et al., 2017). These models can also process sequential data by combining a self-attention mechanism with a positional encoding for each element in the input, along with leveraging an encoder-decoder architecture. Another major improvement is that the computation used in a transformer can easily be parallelized, making it far more efficient to train than basic RNNs.

Transformers are currently an important building block for most state-of-the-art models in computer vision and natural language processing (NLP). One popular model is Bidirectional Encoder Representations from Transformers (BERT), which was developed at Google (Devlin et al., 2018). BERT has been shown in research to achieve impressive results when a pre-trained model is fine-tuned on a variety of common language tasks such as question-answering and sentence continuation.

4 How Deep Learning Is (and Can Be) Used in Credit Unions

4.1 *Deep Learning Models for Consumer Risk Prediction*

Now that we have seen different ways to process sequential data using deep learning models, we can return to the particular task of consumer risk prediction. With these techniques, we can now process a much richer set of input features and format our input to the model not as a fixed-size input vector but as a time-series of events. We could now, for example, more accurately model the scenario in which an individual receives two credit cards in a short period of time as a temporal stream of events as shown below:

- January 1, 2020—Obtained *Visa* credit card with a credit limit of \$2,000
- February 2, 2020—Monthly balance due on all credit cards is \$300
- February 15, 2020—Obtained *Mastercard* with a credit limit of \$10,000
- February 25, 2020—Obtained *Amex* card with a credit limit of \$50,000
- March 1, 2020—Monthly balance due on all cards is \$7,000

By processing the above temporal stream of events, a sequential deep learning model could learn to recognize potentially risky patterns, such as obtaining two credit cards within a ten-day period, and ultimately make more accurate predictions.

Recent papers in the field (Wang et al., 2018) take this a step further and create a dense embedding for each event to further increase the model's complexity. In the context of deep learning, a dense embedding is just a way of mapping a categorical input to a fixed-length vector. One might use, for example, a dense embedding for the input feature of credit card type (Visa, MasterCard, Amex, etc.) to allow the model to have more expressive power. Figure 4 shows an example of using a dense embedding followed by a recurrent architecture.

These state-of-the-art methods for sequential modeling have been implemented on real-world financial datasets with impressive results. One such example is that of Wang et al.'s (2018), which applies these techniques on a dataset of 100,000 borrowers from a lending platform in China. Specifically, this paper investigates the credit risk prediction task using LSTMs, attention mechanisms, and bi-directional LSTMs (which can process information in both the forward and backward temporal dimensions). The author found that, once again, deep learning models outperform traditional machine learning models, with the bi-directional LSTM performing the best overall.

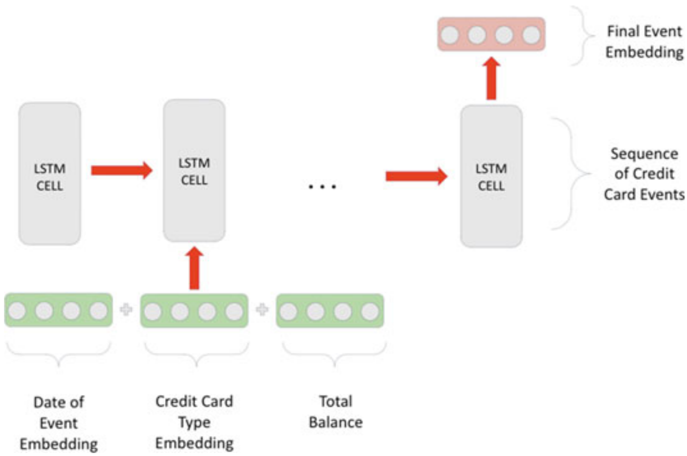


Fig. 4 Example of how a dense embedding can be used in conjunction with a recurrent neural network (RNN) (*Note* Dense tensors are first constructed from categorical variables [credit card type, month of the year] and concatenated together to form the input for one timestep in the sequence)

The main takeaway is that by leveraging deep learning models, an institution can make smarter choices about when to lend credit to consumers and perform more efficiently from a long-term perspective. Furthermore, these predictions can be improved when the inherently sequential nature of consumer data is leveraged, such as by RNN or Transformer model architectures.

4.2 Deep Learning Models for Financial Forecasting

The global financial industry is in a constant state of flux; from new regulations to fluctuating interest rates, any institution that desires to serve its clients as best as possible should be able to identify and adapt to changing conditions. As new credit unions are more vulnerable to changes in the financial environment, it is critical that they anticipate any major structural changes. For example, if there is a rapid decline in equity markets, contagion concerns from an overseas financial crisis, or an unexpected adverse macroeconomic event(s) (inflation, recession, currency crisis, etc.), it is important that the credit union leadership have an early warning system.

Machine learning and big data are core components of such an early warning system.

While news articles provide valuable insight into the current state of the financial system, it is not feasible for a new credit union, with its relatively small staff, to read the large volume of articles about related topics generated daily. As such, a large amount of work focuses on using deep learning models to automatically identify upcoming important financial events. In particular, much research has focused on predicting upcoming distress events, as reacting too slowly to those could lead to severe negative consequences for an institution, especially a new credit union.

However, processing a large corpus of text data and using it to make predictions is not a trivial task. In addition to the scale of the data, much of this type of work is known as unsupervised learning, which operates without a ground truth label for what the text represents. This type of learning contrasts the credit lending scenario described previously, in which we were given the ground truth labels, of whether or not an individual ended up repaying their loan in the training data.

When processing natural language in deep learning, the general approach is first to learn a dense embedding for each word. The training procedure to do so involves trying to predict the next word given a sliding window of the previous n words, similar to the approach used in the popular Word2Vec formulation (Mikolov et al., 2013). This process helps the model understand basic grammar and compositional rules and the context in which each word might be used. Such an approach can extend to learning an embedding of an entire sentence using a two-stage learning process (Rönnqvist & Sarlin, 2016). The first stage exclusively uses the unsupervised approach to generate word embeddings and is given a fixed set of entities (i.e., bank names) to look for in each sentence. Then, the second stage represents each sentence as a fixed-length vector (by aggregating the dense embedding of all the words in the sentence) along with having a ground truth label about whether that sentence related the financial entity to a fixed set of distress events (using a temporal window).

For example, given the sentence "Investors brace themselves for a bailout of Bank XYZ" from a September 2020 article, the event signal would be Bailout of Bank XYZ on Dec 2020 (Rönnqvist & Sarlin, 2016). An illustration of this two-stage process can be seen in Fig. 5.

However, in the aforementioned paper, the authors rightly note that looking at news at the individual bank level is likely to be quite noisy. As such, they aggregate results at a state and then a country level (for various

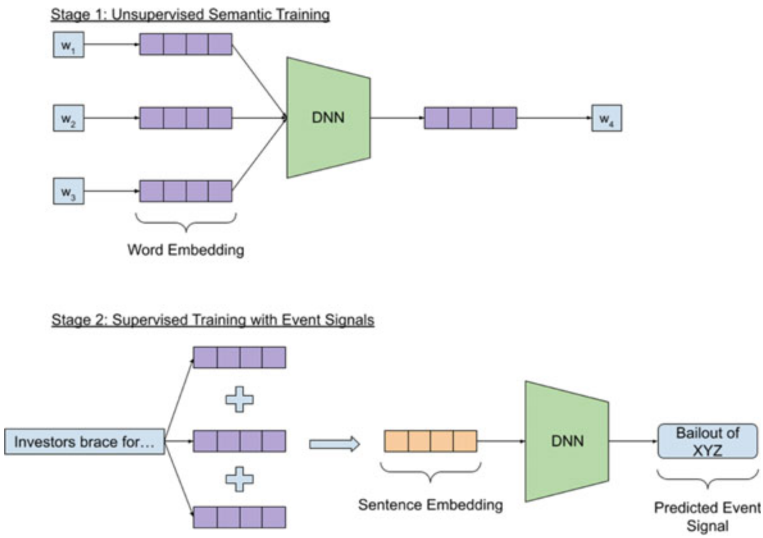


Fig. 5 Two-stage processing for associating words and sentences with an event signal (*Note* In the first stage, basic sentence and word contexts are learned using an unsupervised approach of predicting the next word in a sliding window. In the second stage, supervised training is used to associate a sentence-level embedding with a specified event signal)

countries in Europe) to reduce the variance of their results. The concrete text dataset used in the paper spans a seven-year window from 2007 to 2014, consists of 6.6M articles and uses several large European banks as entities. This window was particularly chosen to encompass major financial events such as the 2008 Financial Crisis. Overall, this type of deep learning model was able to correctly predict a general increase in the level of bank stress leading up to a peak in September 2008 when the financial crisis began. Furthermore, the model was also able to identify individual bailout events for major European banks and predict financial stress levels for several European countries, all in an automated way.

Another common use case for processing financial news data is to predict the movement of stock prices which can ultimately serve as an indicator of the overall market health. Specifically, leveraging sequential deep learning models (such as the previously mentioned BERT model) has been shown to achieve state-of-the-art results on stock movement prediction tasks. In one approach (Chen, 2021), the author demonstrates that by fine-tuning a pre-trained BERT model on headlines from *Bloomberg News*, they can outperform traditional deep learning models (such as LSTMs) on stock market trading simulations. Such results are strong evidence for the ability of transformer models to adapt to new datasets and tasks, which has been observed numerous times in deep learning literature.

Ideally, such modeling approaches could be adapted to Canadian financial news to help provide a "stress index" of the current financial system. By extrapolating the model to predict future events, a financial institution could make adjustments as necessary to weather the course of an economic downturn. More so, in the case of Ontario-based Islamic financial institutions, such unsupervised learning can also be applied at the nexus of Islamic scholarship and modern financial engineering. New financial products are continually being brought to market—whether new reward systems for credit cards or exotic compound derivatives—Islamic scholars must provide guidance on the compatibility of these new developments with the tenants of the Islamic faith. Big data can help predict such compatibility assessment, thus saving resources and time for ICUC management.

5 Conclusions

The ICUC is a unique initiative by any measure; upon its completion, it will be the first deposit-taking institution in Canada to be fully compatible with the tenants of Islam. In no small part, this is due to the power of big data. Without these sophisticated tools to manage risk, convincing regulators and wooing prospective investors would be difficult or even impossible in some cases.

Bringing to bear the latest insights from big data instills confidence that the most advanced tools were used by the modeling team. More so, the models that big data underpins allow both a "telescope" and "microscope" approach. A telescope approach means seeing the big picture or long-term forecast; for example, given a sample of hypothetical loan applicants, we can predict how many loans will be approved, the composition of the portfolio, and the amount of non-performing loans. A microscope approach means we can zoom in and see specific nodes of interest and their impact within a decision tree or process; for example, if one of the features for loan approval is changed, how will that affect the credit union's overall loan portfolio. To receive approval in Ontario for a new credit union, prospective incorporators must meet a high standard. Big data plays a critical role in determining a credit union's feasibility and assisting any future management team in achieving maximum value for its members.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2014). *Neural machine translation by jointly learning to align and translate*. ArXiv. https://doi.org/10.48550/arXiv.1409.0473

Bank for International Settlements. (2017). *Basel III: International regulatory framework for banks*. https://www.bis.org/bcbs/basel3.htm

Chapra, M. U. (2011). *The global financial crisis: Can Islamic finance help? In Islamic economics and finance* (pp. 135–142). Palgrave Macmillan.

Chen, Q. (2021). *Stock movement prediction with financial news using contextualized embedding from BERT*. ArXiv. https://doi.org/10.48550/arXiv.2107.08721

Credit Unions and Caisses Populaires Act, 1994 *(On)* (Can).

Covington, P., Adams, J., & Sargin, E. (2016, September). Deep neural networks for YouTube recommendations. In *Proceedings of the 10th ACM conference on recommender systems* (pp. 191–198). https://research.google/pubs/pub45530/

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805. https://doi.org/10.48550/arXiv.1810.04805

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Machine learning proceedings 1992*. Morgan Kaufmann. https://dl.acm.org/doi/10.5555/141975.142034

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119). https://doi.org/10.48550/arXiv.1310.4546

Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L. J. (2019). A review of deep learning based speech synthesis. *Applied Sciences, 9*(19), 4050. https://doi.org/10.3390/app9194050

Olah, C. (2015, August 27). *Understanding LSTM Networks*. Colah's Blog. http://colah.github.io/posts/2015-08-Understanding-LSTMs/

Ontario Co-operative Association. (2018). *Financial Cooperatives*. https://ontario.coop/financial-co-operatives

Rönnqvist, S., & Sarlin, P. (2016). Bank distress in the news: Describing events through deep learning. *Neurocomputing, 264*, 57–70. https://doi.org/10.1016/j.neucom.2016.12.11

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(6088), 533–536. https://doi.org/10.1038/323533a0

Santos, C. F. G. D., & Papa, J. P. (2022). Avoiding overfitting: A survey on regularization methods for convolutional neural networks. *ACM Computing Surveys (CSUR)*. https://doi.org/10.1145/3510413

Singh, S. P., Kumar, A., Darbari, H., Singh, L., Rastogi, A., & Jain, S. (2017, July). Machine translation using deep learning: An overview. In *2017 international conference on computer, communications and electronics (comptelix)* (pp. 162–167). IEEE. https://doi.org/10.1109/COMPTELIX.2017.8003957

United States Federal Reverse. (2021, November). *Consumer Credit—G.19*. https://www.federalreserve.gov/releases/g19/current/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (p. 30). https://doi.org/10.48550/arXiv.1706.03762

Wang, C., Han, D., Liu, Q., & Luo, S. (2018). A deep learning approach for credit scoring of peer-to-peer lending using attention mechanism LSTM. *IEEE Access, 7*, 2161–2168. https://doi.org/10.1109/ACCESS.2018.2887138

Zhu, B., Yang, W., Wang, H., & Yuan, Y. (2018, May). A hybrid deep learning model for consumer credit scoring. In *2018 International Conference on Artificial Intelligence and Big Data (ICAIBD)* (pp. 205–208). IEEE. https://doi.org/10.1109/ICAIBD.2018.8396195

Assessing the Carbon Footprint of Cryptoassets: Evidence from a Bivariate VAR Model

Hany Fahmy

1 Introduction

On September 7, 2021, El Salvador became the first country to adopt bitcoin as legal tender.[1] A few days later, on September 24, 2021, the central bank of China, one of the world's largest cryptocurrency markets, declared all virtual currency-related business activities illegal. This declaration effectively banned digital tokens such as bitcoin, the first and most popular cryptocurrency. Governments are not alone in their mixed feelings regarding cryptocurrencies and the blockchain technology that scaffolds them. While blockchain technology has many advantages in various fields such as smart contracts (Kosba et al., 2016), insurance (Gatteschi et al., 2018), and education (Turkanovic et al., 2018), among many others, Bitcoin's energy hunger has recently triggered a heated debate in the academic literature about its massive consumption

[1] Through the text, Bitcoin with a capital "B" represents the network while bitcoin with a small "b" represents units of the network's digital currency.

H. Fahmy (✉)
Royal Roads University, Victoria, BC, Canada
e-mail: hany.fahmy@royalroads.ca

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_11

and its carbon footprint (e.g., De Vries, 2018; Hayes, 2017; Kraus & Tolaymat, 2018; Mora et al., 2018; Stoll et al., 2019; Zade et al., 2019). Despite the inconsistencies and the large variations in the results of previous studies when measuring the electricity consumption of Bitcoin itself and, consequently, its carbon footprint, the importance of this topic is seeing a significant increase among policymakers, regulators, and investors, who are becoming acutely aware of climate change risks, especially after the Paris Agreement (Fahmy, 2022b). This topic is also relevant for stakeholders of the new emerging clean energy sector (Fahmy, 2022a).

The main objective of this chapter is to provide a sensible economic approximation and a meaningful economic forecast of the carbon footprint of cryptoassets following Bitcoin's alarming energy consumption. To this end, we examine the relationship between cryptocurrencies' trading volumes and Bitcoin's energy consumption using a vector autoregression (VAR) framework. More specifically, we use the VAR framework to test the directional Granger causality (Granger, 1969) from cryptocurrencies' trading volumes to Bitcoin's electricity consumption. Furthermore, from the VAR estimates, we conduct an impulse-response analysis to investigate the impact of one standard deviation shock (also known as innovation) in bitcoin's trading volume on the network's energy consumption. Finally, we use the results of our impulse-response analysis to provide an approximation of the carbon footprint of bitcoin, and to compute a sensible five-year economic forecast of the environmental impact of such digital currency.

We contribute to the existing literature in several ways. First, by examining the correlation between the top five cryptocurrencies' trading volumes, as proxies of investors' demand for digital currency, and Bitcoin's electricity consumption, we find that the network's electricity consumption correlates highly with trading volumes. This high correlation is robust for all cryptocurrencies. In the case of bitcoin, the correlation coefficient between the currency's average trading volume and the network's electricity consumption is 0.89. This strong positive correlation suggests a similar co-movement of both variables. Our Granger causality tests confirm this fact. Our results show, as expected, the existence of *one-way* directional causality from cryptocurrencies' trading volume to the network's electricity consumption.

Second, after revealing the main economic driver of Bitcoin's electricity consumption, i.e., bitcoin's trading volume, we conduct an impulse-response analysis within the proposed VAR model to measure the impact of a shock in trading volume on electricity consumption. We find that one standard deviation shock in bitcoin's trading volume has a *persistent* impact on Bitcoin's electricity consumption of about 8.8% per month on average over a period of twelve months. This result, which is robust across all cryptocurrencies under investigation, indicates that shocks do not die out in the long run, i.e., after a few months. This result, in turn, means that the network's electricity consumption increases exponentially following increased demand for cryptoassets.

Third, building on the previous dynamic forecast of the impulse–response analysis, we estimate the bitcoin's carbon footprint using reasonable assumptions about the Bitcoin's share of fossil fuels sources, i.e., coal and natural gas, and their carbon intensity. We find that bitcoin mining produced approximately 43 million metric tons of carbon from using coal and natural gas in 2020. This amount represents 0.14% of the world's total carbon emissions of that year. While this figure might not sound high, a deeper look at the progression of Bitcoin's electricity consumption following a shock in its trading volume (as suggested by our impulse–response analysis) reveals a rather disturbing statistic. We find that Bitcoin's average monthly electricity growth of 8.8%, as predicted by our impulse–response analysis, is equivalent to a 63% growth rate per year. This electricity growth means that over five years, i.e., by the end of 2026, and assuming a constant growth of 63% per year, the total electricity consumption of Bitcoin is expected to generate 492 million metric tons of carbon. This electricity consumption translates to approximately 1.6% of today's total carbon emissions worldwide. Indeed, this is a significant environmental impact.

Finally, despite the axiomatic limitations of our predictions, we believe that they are sensible enough to warrant the attention paid to the negative environmental implications of such digital currencies. It is worth noting that our analysis predicts the carbon footprint of bitcoin only. However, bitcoin accounts for two-thirds of the cryptocurrency's total energy consumption, and understudied cryptocurrencies represent the remaining one-third. Understudied currencies add nearly 50% on top of bitcoin's energy hunger (Gallersdörfer et al., 2020). This statistic means that including the remaining cryptocurrencies and the hundreds of other mineable coins and tokens would further increase the share of energy consumption beyond bitcoins. Therefore, further investigation of the carbon footprints of the crypto sector is essential for regulators and policymakers to understand and mitigate the environmental impacts of cryptoassets and the blockchain technology behind them.

The remainder of the chapter is organized as follows: Sect. 2 reviews the literature; Sect. 3 describes the data used in this study; Sect. 4 introduces the VAR methodology and discusses the causality tests and the impulse–response analysis; Sect. 5 documents the environmental impact of Bitcoin's energy consumption; Finally, Sect. 6 concludes the chapter.

2 Literature Review

A cryptocurrency is a decentralized digital monetary and payment system. Most cryptocurrencies utilize blockchain technology, a publicly shared ledger data technology enforced by a network of computers to record transactions continuously among decentralized nodes (Zheng et al., 2017). A blockchain is a distributed database that is continuously updated and verified by its users. Each added record forms a block of data on the chain and becomes part of a growing list of records. Network members surveil all records through a secure database.

The decentralized nature of blockchain technology allows participants to make joint transactions in a shared digital platform without assigning market power to a platform operator. In this respect, this technology contributes to increase competition, lower barriers to entry, and lower privacy risk (Catalini & Gans, 2016). Blockchain technology enables the transfer of assets and secure recording of transactions, and thus, offers innovative advantages in the supply chain, business, healthcare, and other fields, including climate change. A news report by the United Nations Climate Change (UNCC) suggests that blockchain technology can play a major role in the fight against climate change by improving the system of carbon asset transactions, i.e., improving carbon emission trading, promoting peer-to-peer clean energy trading, and enhancing climate finance flow by developing crowdfunding peer-to-peer financial transactions in support of climate action.[2] Sanderson (2018) studies the connection between blockchain and bond markets and suggests that blockchain technology can be a potential solution to enhance the reporting standards of this market. Harris (2018) documents

[2] More details are found here: https://unfccc.int/news/how-blockchain-technology-could-boost-climate-action.

how blockchains can improve environmental goals consistent with the Paris Agreement.

Despite the previously suggested climate benefits of blockchain technology, another strand of the literature warns about the negative environmental impact of using blockchain technology in creating cryptoassets. All bitcoin and most other cryptocurrencies need an algorithmic validation through the resolution of a cryptographic problem to add blocks to the chain, which results in the creation of new cryptocurrency. New bitcoins are created as a reward for transaction processing work in which users offer their computing power to verify and record payments into the public ledger. Individuals or firms engage in this process, which is also known as "mining," in exchange for the chance to earn newly created blocks of bitcoins (Hayes, 2017). While only the node which solves the problem gets the reward, the cryptographic problem is sent to all computer nodes in the network for algorithmic consensus (Dupont, 2019). This situation generates a context where all other computer nodes consume energy without any reason/reward. In fact, this mining process is so computation-intensive that the Bitcoin network has been estimated to consume as much energy (22 terawatt-hours per year) as Ireland in 2018 (De Vries, 2018). Several studies (e.g., De Vries, 2018; Stoll et al., 2019) suggest that the increasing trend in Bitcoin's energy use could seriously increase global temperature. Some studies even suggest that this increase could reach 2 °C by 2034 (e.g., Mora et al., 2018).

In this chapter, we argue that Bitcoin's energy consumption, which is triggered by the continuous increase in cryptocurrencies' trading volumes, is expected to have a severe negative environmental impact. We intend to test this hypothesis and provide a sensible forecast of these impacts. It is worth noting, however, that it is difficult to measure the actual electricity consumption of Bitcoin for various reasons. First, it is difficult to identify the individual miners who operate on Bitcoin's blockchain. Second, miners use different hardware equipment with various degrees of energy efficiency. Third, hashing facilities vary significantly in terms of how effectively they use electric power.[3] For instance, how much electricity is used by miners for cooling as opposed to just running the machines can vary significantly between facilities. For all these reasons, the electricity consumption of Bitcoin cannot be precisely

[3] A hashing facility, also known as "mining farm," is a physical data center dedicated to cryptocurrency mining activities.

determined by researchers. It can, however, be estimated based on theoretical models that rely on specific assumptions. In the literature on the subject, there are many examples of empirically attempting to analyze the electricity consumption of the Bitcoin network and to compute its environmental footprint (e.g., De Vries, 2018; Hayes, 2017; Kraus & Tolaymat, 2018; Mora et al., 2018; Stoll et al., 2019; Zade et al., 2019). As mentioned above, the previous studies produce significantly inconsistent estimates that lead to different assessments of the network's carbon footprint. This inconsistency is due to the different approaches and assumptions used by the authors in their calculations. However, aside from these variations, the previous studies agree that the increasing trend in Bitcoin's energy use has negative environmental implications that warrant further investigations.

3 Data Description

In this chapter, we use the Cambridge Bitcoin Electricity Consumption Index (CBECI) as an estimate of the Bitcoin network's daily electricity load. The monthly CBECI, which is measured in terawatt-hours (TWh), is a bottom-up economic model that provides a best-guess estimate of Bitcoin's actual electricity consumption within the boundaries of a hypothetical lower bound (floor) and a hypothetical upper bound (ceiling) estimate (https://ccaf.io/cbeci/index/methodology). The lower bound estimate is based on the best-case assumption that all miners always use the most energy-efficient equipment available on the market. The upper bound estimate is based on the worst-case assumption that all miners always use the least energy-efficient hardware available on the market. The best-guess estimate, or the CBECI, is based on the more realistic assumption that miners use a basket of profitable hardware rather than a single model. The reason we choose the CBECI approach to compute Bitcoin's electricity consumption is twofold. First, the CBECI model was designed after carefully reviewing the various methodologies and practices used in the literature. Second, except for De Vries' (2018) Bitcoin Energy Consumption Index, which approaches energy consumption from an economic perspective founded on the relation between miners' incomes

Table 1 Summary statistics

Stats	*CBECI (TWh)*	$ATV_{bitcoin}$	$ATV_{ethereum}$	ATV_{ripple}	$ATV_{stellar}$	$ATV_{litecoin}$
Mean	3.8	16.5	8	1.9	0.3	1.8
Maximum	10.3	81	48.7	16.1	2.6	9.2
Minimum	0.2	0	0	0	0	0
S.D	2.9	19.2	10.5	3	0.6	2.2
N	77	77	77	77	77	77

CBECI is the monthly Cambridge Bitcoin Electricity Consumption Index measured in terawatt-hours (TWh)
ATV_j is the monthly average trading volume of cryptocurrency j measured in billions of US dollars
S.D. is the standard deviation
N is the number of observations. The sample period is between August 2015 and December 2021

and costs, the CBECI is the only live index tracking Bitcoin's electricity load and consumption in real-time.[4]

For data on cryptocurrencies, we use daily trading volume data from August 2015 to December 2021 on the five leading cryptocurrencies: Bitcoin, Ethereum, Ripple, Stellar, and Litecoin. These five currencies represent more than 78% of the overall cryptocurrency market and attract more than 82% of the 24-h trade volume (Ji et al., 2019). All daily trading volume series are converted into monthly averages over the analysis period to be consistent with the monthly frequency of the CBECI that is used as a proxy for the cryptocurrency energy consumption. The monthly average trading volumes of the previous currencies are denoted in the text by $ATV_{bitcoin}$, $ATV_{ethereum}$, ATV_{ripple}, $ATV_{stellar}$, and $ATV_{litecoin}$, respectively, and are measured in billions of U.S. dollars. The data is sourced from the Coinmarketcap website (https://coinmarketcap.com). The starting point of the analysis, August 2015, is dictated by data availability for cryptocurrencies. Table 1 gives a summary statistic of the previous variables.

Figure 1 plots the monthly CBECI in TWh and the average trading volume of Bitcoin, Ethereum, Ripple, Stellar, and Litecoin in tens of billions of U.S. dollars over the analysis period (August 2015–December 2021). The average trading volumes of the five cryptocurrencies behave similarly. Bitcoin's average trading volume correlates highly with Ethereum's average trading volume as the correlation coefficient

[4] More details on the methodology and its limitations can be found here: https://digiconomist.net/bitcoin-energy-consumption.

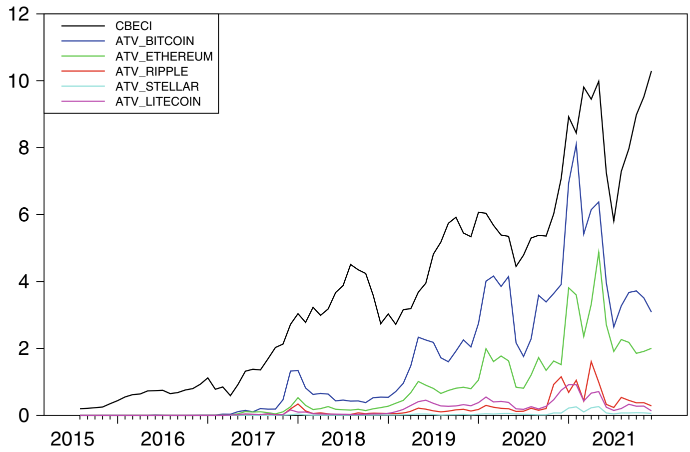


Fig. 1 Monthly Cambridge Bitcoin Electricity Consumption Index (BCECI) and the monthly average trading volume (ATV) of Bitcoin, Ethereum, Ripple, Stellar, and Litecoin (*Note* Monthly Cambridge Bitcoin Electricity Consumption Index [BCECI] measured in terawatt-hours [TWh] and monthly average trading volume [ATV] of cryptocurrencies measured in tens of billions of US dollars between August 2015 and December 2021)

between the two series is 0.96. In fact, all pairwise correlation coefficients between the average trading volumes of all five cryptocurrencies are very high and significant at the 1% level of significance, as shown from the test statistics of Ljung and Box's (1978) Q test of no cross-correlation, which are displayed in brackets beneath the cross-correlations values in the correlation matrix in Table 2. Also notable is the very high and significant correlation between the CBECI and the trading volumes of all cryptocurrencies. This preliminary investigation justifies the similar behavior of these series and points to the existence of causal relation between cryptocurrencies' energy consumption and their trading volume. We expect, of course, one-way directional causality from cryptocurrencies' trading volume to energy consumption and not the other way round. As we demonstrate below, the Granger causality tests within our proposed VAR framework confirm this fact.

4 Empirical Methodology

We use a bivariate vector autoregression (VAR) framework to investigate the relationship between each cryptocurrency's trading volume and

Table 2 Cross-correlations between the Cambridge Bitcoin Electricity Consumption Index (CBECI) and the average trading volumes of the top five cryptocurrencies

Variables	*CBECI*	$ATV_{bitcoin}$	$ATV_{ethereum}$	ATV_{ripple}	$ATV_{stellar}$	$ATV_{litecoin}$
CBECI	1					
$ATV_{bitcoin}$	0.89*** (61.95)	1				
$ATV_{ethereum}$	0.87*** (60.57)	0.96*** (72.95)	1			
ATV_{ripple}	0.73*** (41.67)	0.82*** (53.70)	0.82*** (52.83)	1		
$ATV_{stellar}$	0.75*** (44.16)	0.88*** (62.32)	0.92*** (66.82)	0.87*** (60.21)	1	
$ATV_{litecoin}$	0.77*** (47.48)	0.93*** (67.66)	0.86*** (58.18)	0.82*** (53.65)	0.84*** (56.52)	1

The figures between brackets underneath the pairwise correlation coefficients are the test statistics of Ljung and Box's (1978) robust Q test of zero cross-correlation

***, **, and * denote a test statistic is statistically significant at the 1, 5, and 10% level of significance, respectively

energy consumption. To reduce any unwanted variability (heteroskedasticity) in the data, we transform all series into their natural logarithms. To simplify the notation, let $Y = \ln(\text{CBECI})$, $X^b = \ln(ATV_{bitcoin})$, $X^e = \ln(ATV_{ethereum})$, $X^r = \ln\left(ATV_{ripple}\right)$, $X^s = \ln(ATV_{stellar})$, and $X^l = \ln(ATV_{litecoin})$. In a VAR framework, each variable depends on the lagged values of all the variables in the system. For instance, in a VAR system of order p that consists of two variables Y and X^j, for any cryptocurrency j, i.e., a bivariate VAR of order p, Y depends on its p lags and the p lags of X^j, and similarly, X^j depends on its own p lags and the p lags of Y; that is, a bivariate VAR system of order p that consists of the Bitcoin electricity consumption and cryptocurrency j's trading volume is defined as:

$$Y_t = \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \theta_1 X^j_{t-1} + \ldots + \theta_p X^j_{t-p} + \varepsilon_t,$$

$$X^j_t = \psi_1 X^j_{t-1} + \ldots + \psi_p X^j_{t-p} + \lambda_1 Y_{t-1} + \ldots + \lambda_p Y_{t-p} + \nu_t, \quad (1)$$

for $t = 1, \ldots, N$, where ϕ_i, θ_i, ψ_i, λ_i, for $i = 1, \ldots, p$, are system parameters, and ε_t, ν_t are error terms that are assumed to be independent and identically distributed with a zero mean and a constant variance.[5] One of the advantages of using the previous VAR framework is that it allows us to test the Granger directional causality (i.e., whether the lags of one variable help to explain the current value of the other variable) and describe the dynamic of the data (i.e., the response of one variable to a one standard deviation shock in another). For instance, from the first equation in the previous system, we can test whether the lags of the trading volume of cryptocurrency X^j are jointly causing the current value of the Bitcoin network's electricity consumption by simply testing the null hypothesis $H_{01} : \theta_1 = \ldots = \theta_p = 0$. Similarly, the other directional causality can be tested by testing the null hypothesis $H_{02} : \lambda_1 = \ldots = \lambda_p = 0$ from the second equation in the system. Rejecting H_{01} and accepting H_{02} is indicative of a one-way directional causality from cryptocurrency trading volume to electricity consumption.

Another advantage of the rich data structure of the VAR system is that the model dynamic can be studied using impulse–response functions. A linear impulse–response function within a VAR framework shows the impact of one standard deviation shock in one variable on another variable's current and future values. Using the Cholesky decomposition, we compute the orthogonalized impulse–response functions. We choose these functions over the conventional linear impulse–response functions because, unlike the linear type, this type forces the innovation in one time series to have no contemporaneous effect on the other series. An orthogonalized impulse–response function requires an ordering of the variables such that the first variable in the ordering is not contemporaneously affected by shocks to the second variable. However, the first variable is the only one with a potential immediate impact on the second. This setup, which is consistent with the (Granger) causality from cryptocurrency j's trading volume, X^j, to Bitcoin's electricity consumption, Y, is particularly suitable here as it allows us to easily trace the impact of a shock in trading volume (first variable in the ordering) on electricity consumption (second variable in the ordering). The results of this analysis will subsequently allow us to assess the environmental footprint of cryptocurrencies' energy consumption using sensible assumptions.

We begin the analysis by selecting an appropriate lag length p for the VAR system in Eq. (1). Determining the proper lag length in a VAR

[5] The order p in each equation refers to the number of lags of each variable.

system depends on the purpose of fitting such a model. If the objective is to test the directional (Granger) causality among the variables, then using a selection criterion like the Akaike information criterion, or AIC for short (Akaike, 1974), is advisable. If, however, the objective is to study the dynamics of the data using impulse–response functions, then selecting a reasonable lag length for the data (such as $p = 4$ for quarterly and $p = 12$ for monthly data) is recommended. It is also recommended to ensure that the maximum lag length is reasonable for the model's size, i.e., the number of observations. In the present analysis, we will use the AIC to determine the lag length of the VAR model to test the directional causality from cryptocurrencies' trading volume to energy consumption. As for the dynamic analysis, we will use a lag length $p = 12$ to study the impact of one standard deviation shock in trading volume on energy consumption. This lag length is reasonable here since we have a total number of $N = 77$ monthly observations.

4.1 Causality Tests

We begin with the directional causality analysis. Guided by the AIC, which is minimized at a lag length of order 1 for the bivariate VAR system for Y and X^j for each j, we fit the following VAR model of order $p = 1$ to each pair of variables for each j:

$$Y_t = \phi_1 Y_{t-1} + \theta_1 X^j_{t-1} + \varepsilon_t,$$

$$X^j_t = \psi_1 X^j_{t-1} + \lambda_1 Y_{t-1} + \nu_t, \tag{2}$$

where $t = 1, \ldots, N$, and everything else is defined as before.

The estimation results suggest that the bivariate VAR model for each pair Y and X^j, for every j, fits well. Table 3 documents the results. Consider the Y-X^b pair, for instance. The adjusted coefficient of determination (adjusted R^2) is 0.982 for the Y equation and 0.980 for the X^b equation. Similar values are reported for the other four pairs of equations. These values are remarkably high and demonstrate the goodness of fit of the bivariate VAR models. The standard error of each equation measures how different the dependent variable's predicted values are from the actual values. Smaller standard error values are better because they imply less dispersion about the regression line. This smaller standard error, in turn, means a tighter fitting model. All equations' standard error values obtained are small, which is another indication of the goodness of fit ofall fitted bivariate VAR models. The standard error of the equation expressed as a percentage of the mean of the dependent variable also confirms that each equation fits very well. Finally, we test the goodness of fit for each equation by adding a constant term to each equation in the VAR system and testing the null hypothesis that all coefficients (except the constant term) of the explanatory variables on the right-hand side are zeroes. Judging by the F statistic of this test, we reject this null hypothesis at the 1% level of significance for each equation and conclude that the overall fit of each equation is significant.

Next, we perform Granger causality tests on each bivariate VAR system ($\mathrm{Y} - \mathrm{X}^j$ for each j) and report the results in Table 4. Judging by the F statistic of the tests, the null hypothesis that the trading volume of cryptocurrency j, X_{t-1}^j, does not Granger cause the Bitcoin network's electricity consumption, Y_t, is rejected at the 1% level of significance for all five cryptocurrencies, whereas the null hypothesis that electricity consumption does not cause trading volume is accepted. Therefore, the previous tests confirm the existence of one-way directional (Granger) causality from cryptocurrency trading volume to energy consumption.

4.2 Impulse–Response Analysis

Finally, for each pair ($\mathrm{Y} - \mathrm{X}^j$), we compute the orthogonalized impulse–response functions over a period of a twelve-month time horizon. For each shock, we construct 95% Monte Carlo bands to confirm the statistical significance of its impact. Figure 2 depicts the response of the natural logarithm of Bitcoin's electricity consumption, Y, to one standard deviation shock in the natural logarithm of bitcoin's trading volume, X^b, over twelve months. The immediate impact of the shock on electricity consumption is about 5%. The shock then surges to about 6–10% monthly magnitude (or an average of 8.8% per month) and *persists* over the entire year. Judging by the orthogonalized impulse–response functions for the remaining four cryptocurrencies in Fig. 3, the average persistence of 8% impact on electricity consumption due to one standard deviation shock in trading volume seems to be a common impulse–response for all cryptocurrencies.

Table 3 Goodness of fit of the bivariate vector autoregression (VAR) models

Stats	Y	X^b	Y	X^e	Y	X^r	Y	X^s	Y	X^l
Adjusted R^2	0.982	0.980	0.980	0.970	0.983	0.956	0.982	0.947	0.982	0.963
S.E. equation	0.138	0.357	0.144	0.543	0.135	0.685	0.139	0.883	0.137	0.543
Mean dependent	0.938	−1.003	0.938	−2.180	0.938	−3.845	0.938	−6.024	0.938	−3.468
S.E. (%)	0.147	−0.356	0.154	−0.249	0.144	−0.178	0.148	−0.147	0.146	−0.157
F statistic	2092.2***	1751.4***	1931.2***	1222.7***	2197.6***	827.2***	2069.4***	681.1***	2106.8***	995.5***

Adjusted R^2 is the adjusted coefficient of determination. S.E. denote standard error. S.E.(%) is the standard error of each equation expressed as a percentage of the mean of the dependent variable. *F* is the statistic of the overall *F* test of significance
***, **, and * denote a test statistic is statistically significant at the 1%, 5%, and 10% level of significance, respectively

Table 4 Results of Granger Causality tests

VAR system for currency j: $Y_t = \phi_1 Y_{t-1} + \theta_1 X^j_{t-1} + \varepsilon_t$ $X^j_t = \psi_1 X^j_{t-1} + \lambda_1 Y_{t-1} + \nu_t$	*Null hypothesis:* $H_{01}: \theta_1 = 0$ *(*X^j_{t-1} *does not Granger cause* Y_t*)* $H_{02}: \lambda_1 = 0$ *(*Y_{t-1} *does not Granger cause* X^j_t*)*	$F(n_1, n_2)$	*p-value*
$j = b$	Reject H_{01}	$F(1, 74) = 8.79^{***}$	0.004
	Accept H_{02}	$F(1, 74) = 0.03$	0.862
$j = e$	Reject H_{01}	$F(1, 74) = 10.91^{***}$	0.001
	Accept H_{02}	$F(1, 74) = 0.57$	0.452
$j = r$	Reject H_{01}	$F(1, 74) = 8.463^{***}$	0.004
	Accept H_{02}	$F(1, 74) = 0.037$	0.847
$j = s$	Reject H_{01}	$F(1, 74) = 10.447^{***}$	0.002
	Accept H_{02}	$F(1, 74) = 0.064$	0.801
$j = l$	Reject H_{01}	$F(1, 74) = 8.462^{***}$	0.005
	Accept H_{02}	$F(1, 74) = 0.085$	0.771

$F(n_1, n_2)$ is the F statistic of the test, where n_1 and n_2 are the degrees of freedom of the numerator and denominator of the statistic, respectively. ***, **, and * denote a test statistic is statistically significant at the 1%, 5%, and 10% level of significance, respectively

5 Environmental Impact of Cryptoassets

We have established the existence of a one-way directional causality between cryptocurrency trading and electricity consumption. Furthermore, we have demonstrated that a trading shock has a persistent impact of about 8.8% on average per month on electricity consumption. In this section, we intend to use the previous results to assess the environmental footprint of Bitcoin. In this context, it is essential to remember that Bitcoin's electricity consumption refers to the total amount of electricity used by the Bitcoin mining process. However, what ultimately matters for the environment is not the total amount of electricity consumption; instead, it is the *carbon intensity of the energy source* used to generate this amount of electricity. For instance, one TWh of coal station-generated electricity has a significantly worse carbon footprint than one TWh of wind- or solar farm-generated electricity. Thus, to assess the environmental footprint of cryptoassets, we should investigate the energy sources used in their mining. So, what energy sources do Bitcoin miners use? Unfortunately, we do not have accurate data on either the type or the

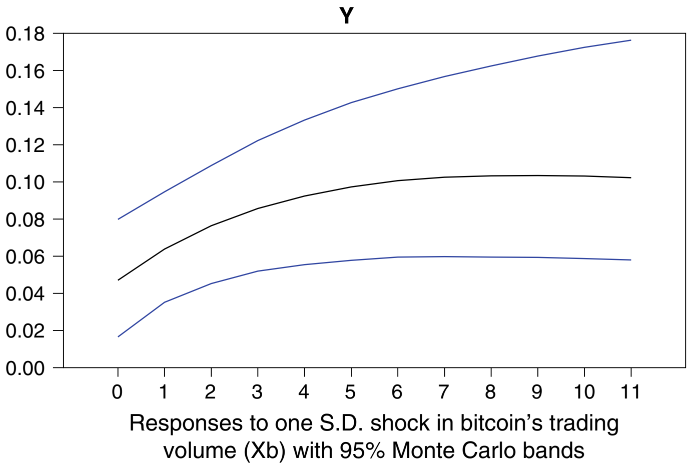


Fig. 2 Orthogonalized response of Bitcoin's electricity consumption to a one standard deviation impulse (shock) of Bitcoin's trading volume (*Note* Orthogonalized response and impulse shock in the natural logarithm. Blue lines represent 95% Monte Carlo bands)

number of energy sources used by Bitcoin miners. What we do know, however, is that Bitcoin miners use a wide variety of power sources, including coal, natural gas, hydroelectric power, nuclear power, and oil. In a recent industry survey, Blandin et al. (2020) find that hydroelectric power, coal, and natural gas are the dominating power sources. The survey also finds that Bitcoin miners use renewable power sources (wind, solar, and geothermal).

Against the previous backdrop, it is clear that until better data on the network's power mix become available, the only way to assess the environmental footprint of cryptoassets is by making assertions about the type and the quantity of the network's source power mix. These assertions, however, should be reasonable to avoid radical misleading predictions. As a guide, we will use the U.S. power mix distribution to compute the total electricity generated from polluting power sources. Although China's share of the global hash rate, which measures the total computational power that is dedicated to mining, amounted to roughly 76% in September 2019, the country's recent ban on all cryptocurrencies' related activities has significantly increased the U.S.'s share of global Bitcoin

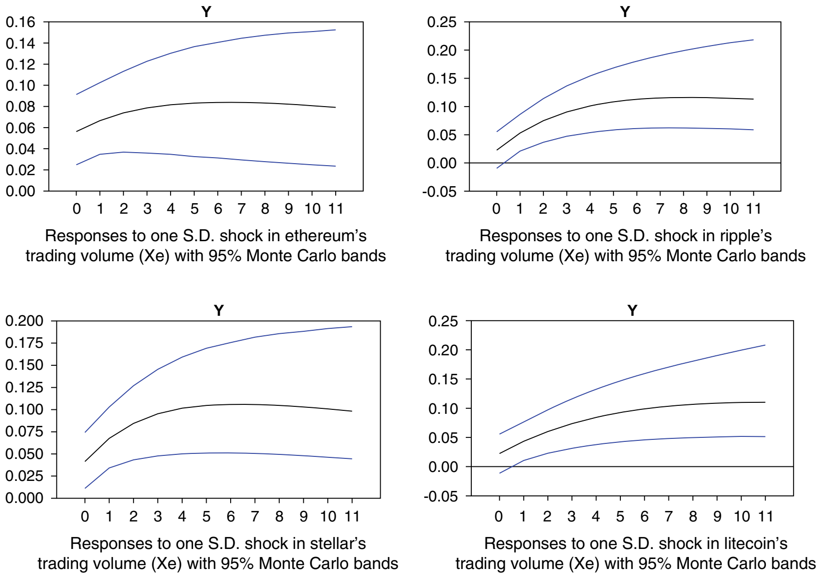


Fig. 3 Orthogonalized response of Bitcoin's electricity consumption to one standard deviation impulse (shock) of the monthly average trading volume (ATV) of Ethereum, Ripple, Stellar, and Litecoin (*Note* Orthogonalized response and impulse shock in the natural logarithm. Blue lines represent 95% Monte Carlo bands)

hash rate from 4.1% in 2019 to 35.4% in 2021, per Table 5. Thus, using the U.S. data as a guide for the network's source power mix is sensible. More specifically, we use the power mix distribution of total electricity net generation for all U.S. sectors from the most recent Monthly Energy Review report issued by the U.S. Energy Information Administration (EIA) in December 2021 (https://www.eia.gov/totalenergy/data/monthly/pdf/sec7.pdf). Table 6, which is reproduced from Table 7.2a of the 2021 EIA's Monthly Energy Review report, documents the last five years of complete annual data on the total electricity net generation from fossil fuels (coal, petroleum, natural gas, and other gases), nuclear electric power, and renewable energy (hydroelectric power, biomass, geothermal,

Table 5 Global hash rate by country

Country	*Share of global hash rate as of September 2019 (%)*	*Share of global hash rate as of August 2021 (%)*
U.S	4.1	35.4
Kazakhstan	1.4	18.1
Other	6.1	13.5
Russia	5.9	11.2
Canada	1.1	9.6
Malaysia	3.3	4.6
Germany	0.9	4.5
Iran	1.7	3.1
China	75.5	0.0

Source This data is obtained from the mining map of the Cambridge Bitcoin Electricity Consumption Index (CBECI), which is housed by the University of Cambridge: https://ccaf.io/cbeci/mining_map

solar, and wind) in million kilowatt-hours (kWh).[6] Taking the average net generation of each source relative to the total average electricity generation, we find that, on average, 63% of total electricity generation comes from fossil fuels (mainly 26% from coal and 36% from natural gas), 20% is coming from nuclear electric power, and 17% is coming from renewable energy sources as documented in the third column of Table 7.

The next step is to find an estimate of the carbon emissions of fossil fuels—the polluting source of electricity generation. Again, we will revert to the EIA's recent U.S. data on carbon emissions by fuel type in 2020 (https://www.eia.gov/tools/faqs/faq.php?id=74&t=11). According to the EIA, the 2020 carbon emissions in million metric tons ($MtCO_2$) per million kWh for coal, petroleum, and natural gas are 0.00101219, 0.000952032, and 0.000410713, respectively.

Finally, we compute the most recent annual Bitcoin electricity consumption in 2021 by simply aggregating the monthly data of the CBECI in 2021. This data aggregation yields approximately a total electricity consumption of 104 TWh (or 104,000 million kWh). If we entertain the previous U.S. energy mix distribution and carbon emissions

[6] The report has only 9 months of data in 2021. For that reason, this year is excluded from the analysis.

Table 6 Total electricity net generation from all sources and sectors in the United States

Year	*Fossil Fuels*				*Nuclear Electric Power*
	Coal	*Petroleum*	*Natural Gas*	*Other Gases*	
2016	1,239,148.654	24,204.806	1,378,306.934	12,807.432	805,693.948
2017	1,205,835.276	21,389.945	1,296,442.491	12,468.967	804,949.635
2018	1,149,487.339	25,225.618	1,469,132.682	13,462.749	807,084.477
2019	964,956.812	18,340.743	1,585,814.174	12,591.094	809,409.262
2020	773,392.897	17,341.014	1,624,049.997	11,818.478	789,878.863
Average	1,066,564.2	21,300.425	1,470,749.26	12,629.744	803,403.237

Year	*Renewable Energy*					*Totals*
	Hydro-electric Power[a]	*Biomass*[b]	*Geo-thermal*	*Solar*	*Wind*	*Total*[c]
2016	261,126.026	62,760.458	15,825.807	36,054.121	226,992.562	4,076,674.984
2017	293,838.382	62,733.412	15,926.774	53,286.865	254,302.695	4,034,270.559
2018	286,619.45	61,831.924	15,967.134	63,825.315	272,667.454	4,178,277.344
2019	282,612.987	57,506.953	15,472.717	71,936.822	295,882.484	4,127,855.214
2020	279,952.69	54,703.01	15,889.697	89,198.715	337,938.049	4,007,018.595
Average	280,829.907	59,907.1514	15,816.426	62,860.368	277,556.649	4,084,819.34

[a]This category includes both the hydroelectric pumped storage and the conventional hydroelectric power. The hydroelectric pumped storage is the pumped storage facility production minus energy used for pumping

[b]This includes wood, wood-derived fuels, municipal solid waste from biogenic sources, landfill gas, sludge waste, agricultural by-products, and other biomass

[c]The total electricity generation figures is slightly larger than the exact total of all categories since it includes other minor sources that are not listed in the table

statistics, then the carbon emissions from the 26% share of coal are:

$$0.26 \times 104,000\text{millionkWh} \times \frac{0.00101219\text{millionMtCO}_2}{1\text{millionkWh}} \approx 27.4\text{millionMtCO}_2, \tag{3}$$

and the carbon emissions from the 36% share of natural gas are

$$0.36 \times 104,000\text{millionkWh} \times \frac{0.000410713\text{millionMtCO}_2}{1\text{millionkWh}} \approx 15.4\text{millionMtCO}_2. \tag{4}$$

Thus, the total carbon emissions from the two main fossil fuel sources (coal and natural gas) used to generate the electricity needed for Bitcoin

Table 7 Average proportion of total electricity net generation from fossil fuels, nuclear power, and renewable energy between 2016 and 2020 in the United States

Type of energy source used for electricity generation	*Average net electricity generation per source between 2016 and 2020 (million kWh)*	*Average proportion of net energy generation per source*
Coal	$a = 1066564.2$	$(a/h) \times 100 = 26\%$
Petroleum	$b = 21300.425$	$(b/h) \times 100 = 0.5\%$
Natural Gas	$c = 1470749.26$	$(c/h) \times 100 = 36\%$
Other Gases	$d = 12629.744$	$(d/h) \times 100 = 0.3\%$
Fossil Fuels	$e = a + b + c + d = 2571243.62$	$(e/h) \times 100 = 63\%$
Nuclear Electric Power	$f = 803403.237$	$(f/h) \times 100 = 20\%$
Renewable Energy	$g = 696970.501$	$(g/h) \times 100 = 17\%$
Total	$h = e + f + g = 4084819.34$	100%

mining are approximately 43 million $MtCO_2$. This estimate is consistent with Stoll et al. (2019)'s figures except for the minor discrepancy in the estimate of Bitcoin's energy consumption. The authors find that Bitcoin's annual electricity consumption adds up to 45.8 TWh in 2018 and the corresponding annual carbon emissions range from 22.0 to 22.9 $MtCO_2$. In our analysis, we use the CBECI model to compute electricity consumption. According to this model, electricity consumption in 2018 was 52.18 TWh. Plugging this value in Eqs. (3) and (4) and adding up the resulting carbon emissions yield a total of 21.4 $MtCO_2$, which is very close to the lower bound of the reported range of Stoll et al. (2019). What is the significance of 22 million tons of carbon emissions in 2018 or 43 million tons of carbon emissions in 2020? Stoll et al. (2019) document that a level of 22 million $MtCO_2$ sits between the levels produced by the nations of Jordan and Sri Lanka. Obviously, 43 million $MtCO_2$ in 2020 is almost double this amount. For a more global perspective, a recent IEA study finds that the global energy-related carbon emissions amounted to around 31,500 million $MtCO_2$ in 2020.[7] Thus, our calculations of

[7] More details are found here: https://www.iea.org/articles/global-energy-review-co2-emissions-in-2020.

Bitcoin's carbon footprint in 2020 account for roughly $\frac{43}{31500} \approx 0.14\%$ of the world's total yearly emissions.

A carbon footprint amounting to 0.14% of the world's total emissions might not seem alarming. However, a deeper look at the progression of electricity consumption following a shock in cryptoassets' trading volume is indeed eye-opening. To demonstrate this, we use our earlier prediction from the impulse–response analysis in Sect. 4 to forecast the impact on electricity consumption following one standard deviation shock inbitcoin's trading volume over a time horizon of 12 months. We began with the CBECI's most recent recorded value in December 2021. The index shows that the total cumulative electricity consumption of Bitcoin (from January to December 2021) is 103.72 TWh. This obtained value is our initial forecasting value, which is recorded in the intersection of the first row and column *a* of Table 7. Next, we transform this initial value to its natural logarithm value since the trading volume, $X^b = \ln(ATV_{bitcoin})$, and the electricity consumption, $Y = \ln(CBECI)$, in the impulse–response analysis are in natural logarithms. We record the forecasted value of Y in column *b* of Table 8. We record the monthly responses of Y to one standard deviation impulse in X^b that is revealed by our previous impulse–response analysis in column *c* of Table 8. In column *d*, we isolate the marginal response (that is, the change in the monthly responses of Y following the shock in X^b). For instance, the first predicted response of Y in month 1 is 4.58%. The total response in two months is 6.18%. Thus, the marginal response of Y in month 2 is simply the difference between 6.18% and 4.58%, which is 1.6%. Next, in column *e*, we grow the initial value of Y in column *b* by the marginal monthly response values in column *d* to obtain monthly forecasted values of Y measured in natural logarithm TWh. Finally, we convert back the value of Y, i.e., the natural logarithm of CBECI, to just CBECI using the exponential (anti-log) function in column *f*. We repeat the previous exercise for every month over the one-year forecast horizon. We find that a shock in bitcoin's trading volume leads to an estimated increase in Bitcoin's electricity consumption from approximately 104 TWh at the beginning of the year to 169 TWh by the end of the year. This increase means that the predicted surge in electricity consumption is about 63% per year. This result is indeed a significant number. If we entertain the oversimplified assumption that this increase in electricity will remain constant over the next five years (provided that cryptocurrency remains appealing to investors and other stakeholders), then the total electricity consumption of Bitcoin by the end of 2026 is expected to be $169 \times (1 + 0.63)^5 = 1196,662$ million kWh. Plugging this value back in Eqs. (3) and (4) yields a total of 492 million metric tons of carbon emissions from coal and natural gas usage as:

$$0.26 \times 1196,662\text{millionkWh} \times \frac{0.00101219\text{millionMtCO}_2}{1\text{millionkWh}} \approx 315\text{millionMtCO}_2 \tag{5}$$

and

$$0.36 \times 1196,662\text{millionkWh} \times \frac{0.000410713\text{millionMtCO}_2}{1\text{millionkWh}} \approx 177\text{millionMtCO}_2 \tag{6}$$

Table 8 One year prediction of Bitcoin's electricity consumption

Column	*a*	$b = \ln(a)$	*c*	$d = \Delta c$	$e = b \times \left(1 + \frac{d}{100}\right)$	$f = \exp\{e\}$
Time horizon (month)	*Forecasted value of CBECI (TWh)*	*Forecasted value of* $Y = \ln(CBECI)$ *(ln TWh)*	*Response of* $Y = \ln(CBECI)$ *to shock in* X^b *(%)*	*Marginal response of Y to shock in* X^b *(%)*	*Forecasted value of Y (ln TWh)*	*Forecasted value of C BECI (TWh)*
1	103.72	4.64	4.585632	4.585632	4.85454601	128.32
2*	128.32	4.85	6.185784	1.600152	4.93509081	139.09
3	139.09	4.94	7.413921	1.228137	4.99570048	147.78
4	147.78	5.00	8.341742	0.927821	5.04205164	154.79
5	154.79	5.04	9.027479	0.685737	5.07662685	160.23
6	160.23	5.08	9.518419	0.49094	5.10155005	164.28
7	164.28	5.10	9.85295	0.334531	5.11861631	167.10
8	167.10	5.12	10.062226	0.209276	5.12932835	168.90
9	168.90	5.13	10.171518	0.109292	5.13493429	169.85
10	169.85	5.13	10.201323	0.029805	5.13646476	170.11
11	170.11	5.14	10.168251	−0.03307	5.13476603	169.82
12	169.82	5.13	10.085753	−0.0825	5.13052995	169.11

This table reads from left to right row wise beginning with the first row. The value 103.72 TWh at the intersection of month 1 and Column *a* is the CBECI's value of total electricity consumption of the Bitcoin network as of December 2021. This is the initial forecast value

*The values starting in month 2 in Column *a* are the one-month lagged values of Column *f* respectively. These results yield an estimated value of Bitcoin's carbon footprint in 2026 of roughly $\frac{492}{31500} \approx 1.6\%$ of the world's total yearly emissions.

6 Concluding Remarks

A prediction that Bitcoin's carbon footprint in 2026 amounts to 1.6% of today's total carbon emissions worldwide is disturbing. Even worse, this analysis does not include the carbon footprint of several factors that could potentially make this figure even larger. First, our analysis does not include the carbon footprint of Bitcoin's entire hardware supply chain from production to delivery. Second, it does not include the carbon footprint from the e-waste that the disposal of older models generates. Third, the analysis does not include other understudied cryptocurrencies, which could add 50% on top of bitcoin's energy hunger (Gallersdörfer et al., 2020).

We acknowledge several limitations of the present analysis. First, we use the CBECI model to estimate Bitcoin's electricity consumption. This model, like any other model, has its limitations. However, the estimates seem reasonable and consistent with other studies (e.g., Stoll et al., 2019). Second, due to data unavailability, we use the power mix distribution of total electricity net generation for all U.S. sectors from the EIA as a proxy for the proportions of fossil fuel sources, i.e., coal and natural gas, that generate the total electricity of Bitcoin. We also use carbon emissions data on these energy sources from the EIA to estimate the carbon footprint of the network. Having said that, we consider the prediction that this chapter presents as a cause for reasonable concern about the environmental impact of Bitcoin. Although our results may differ from the actual carbon emissions figures, we believe that our assumptions are sensible enough to warrant attention to the negative environmental implications of such digital currency.

In closing, despite this study's data and axiomatic limitations, it is clear that a continuous increase in demand for cryptocurrencies, as reflected in the currencies' trading volumes, will create persistent and tremendous growth in the network's energy consumption, and in turn, its carbon footprint. This observation alone, without an actual or predicted figure for carbon emissions, warrants two courses of action. First, regulators and policymakers should design climate mitigation policies that effectively address the environmental impacts of cryptoassets. A carbon tax, for instance, is not an effective policy because it does not eliminate the carbon emissions from mining. However, a policy that forces miners to generate electricity only from green energy sources, e.g., solar, wind, and geothermal, is more effective in eliminating the carbon footprint of this sector. Second, regulators and policymakers should carefully assess the societal costs (including the alarming carbon footprint) and benefits of cryptocurrencies as a decentralized monetary system. Even if bitcoin mining became totally green, one must not ignore the opportunity cost of Bitcoin's tremendous electricity consumption, i.e., the benefit that otherwise could have been obtained from alternative use of Bitcoin's renewable energy sources. Do we need bitcoins as a currency? Do we need to exhaust green energy sources to generate this digital currency? Do the costs of doing so exceed the benefits? These are all interesting future research questions on this critical topic.

Acknowledgements The author would like to thank the editing team and two anonymous reviewers for their helpful comments. This paper is financially supported by a research grant from Royal Roads University.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723.

Blandin, A., Pieters, G., Wu, Y., Eisermann, T., Dek, A., Taylor, S., & Njoki, D. (2020). *3rd global cryptoasset benchmarking study*. Judge Business School, University of Cambridge.

Catalini, C., & Gans, J. S. (2016). *Some simple economics of the blockchain* (NBER Working Paper No. 22952, NBER).

De Vries, A. (2018). Bitcoin's growing energy problem. *Joule, 2*, 801–805.

Dupont, Q. (2019). *Cryptocurrencies and blockchains*. Polity Press.

Fahmy, H. (2022a). Clean energy deserves to be an asset class: A volatility-reward analysis. *Economic Modelling, 106*, 105696.

Fahmy, H. (2022b). The rise in investors' awareness of climate risks after the Paris Agreement and the clean energy-oil-technolog prices nexus. *Energy Economics, 106*, 105738.

Gallersdörfer, U., Klaaßen, L., & Stoll, C. (2020). Energy consumption of cryptocurrencies beyond bitcoin. *Joule, 4*, 1843–1846.

Gatteschi, V., Lamberti, F., Demartini, C., Paranteda, C., & Santamaria, V. (2018). Blockchain and smart contracts for insurance: Is the technology mature enough? *Future Internet, 10*(2), 20.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica, 37*, 424–438.
Harris, A. (2018). A conversation with masterminds in blockchain and climate change. In A. Marke (Ed.), *Transforming climate finance and green investment with blockchains* (Ch. 2., pp. 15–22). Academic Press.
Hayes, A. S. (2017). Cryptocurrency value formation: An empirical study leading to a cost of production model for valuing bitcoin. *Telematics and Informatics, 34*, 1308–1321.
Ji, Q., Bouri. E., Roubaud, D., & Kristoufek, L. (2019). Information interdependence among energy cryptocurrency and major commodity markets. *Energy Economics*, 811042–811055. S0140988319301896 https://doi.org/10.1016/j.eneco.2019.06.005
Kosba, A., Miller, A., Shi, E., Wen, Z., & Papamanthou, C. (2016). Hawk: The blockchain model of cryptography and privacy-preserving smart contracts. In *Proceedings of the 2016 IEEE Symposium on Security and Privacy* (S.P.).
Krause, M. J., & Tolaymat, T. (2018). Quantification of energy and carbon costs for mining cryptocurrencies. *Nature Sustainability, 1*, 711–718.
Ljung, G. M., & Box, G. E. P. (1978). On a measure of a lack of fit in time series models. *Biometrika, 65*(2), 297–303.
Mora, C., Rollins, R. L., Taladay, K., Kantar, M. B., Chock, M. K., Shimada, M., & Franklin, E. C. (2018). Bitcoin emissions alone could push global warming above 2° C. *Nature Climate Change, 8*(11), 931–933.
Sanderson, O. (2018). How to trust green bonds: blockchain, climate, and the institutional bond markets. In A. Marke (Ed.), *Transforming climate finance and green investment with blockchains* (Ch. 20., pp. 273–288). Academic Press.
Stoll, C., Klaaßen, L., & Gallersdörfer, U. (2019). The carbon footprint of bitcoin. *Joule, 3*(1), 1647–1661.
Turkanovic, M., Hölbl, M., Košic, K., Hericko, M., & Kamišalic, A. (2018). EduCTX: A blockchain-based higher education credit platform. *IEEE Access, 6*, 5112–5127.
Zade, M., Myklebost, J., Tzscheutschler, P., & Wagner, U. (2019). Is Bitcoin the only problem? A scenario model for the power demand of blockchains. *Frontiers in Energy Research, 7*, 21.
Zheng, Z., Xie, S., Dai, H., Chen, X., & Wang, H. (2017). An overview of blockchain technology: Architecture, consensus, and future trends. *International Congress of Big Data (BigData Congress), 2017*, 557–564.

A Data-Informed Approach to Financial Literacy Enhancement Using Cognitive and Behavioral Analytics

Prasanta Bhattacharya, Kum Seong Wan, Boon Kiat Quek, Waseem Bak'r Hameed, and Sivanithy Rathananthan

1 Introduction

In the past decade, there has been a remarkable increase in both the relevance of, and interest in, the topic of financial literacy among citizens and governments[1] (Huston, 2010; Lusardi & Mitchell, 2014). The need for government interventions in financial literacy, planning, and well-being has become particularly salient in the aftermath of the COVID-19

[1] https://ec.europa.eu/commission/presscorner/detail/en/ip_22_283.

P. Bhattacharya (✉) · K. S. Wan · B. K. Quek
Institute of High Performance Computing (IHPC), Singapore, Singapore
e-mail: prasantab@ihpc.a-star.edu.sg

K. S. Wan
e-mail: kswan@ihpc.a-star.edu.sg

B. K. Quek
e-mail: quekbk@ihpc.a-star.edu.sg

W. B. Hameed · S. Rathananthan
Institute for Financial Literacy (IFL), Singapore, Singapore
e-mail: waseem_hameed@spi.edu.sg

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8_12

pandemic that has created tremendous financial burdens on households,[2] increased financial distress and anxiety, increased healthcare expenditures for organizations (Barnett et al., 2020), and generated substantial economic uncertainty for the near future.[3] We have also witnessed a rapid transition toward the digitalization of financial literacy interventions, with governments and financial institutions now offering a wide range of online touchpoints for digital literacy as well as financial literacy enhancement (Lyons & Kass-Hanna, 2021). The dual growth of digital literacy and digital finance has made it imperative to develop a systematic understanding of what constitutes financial literacy enhancement, and how we can harness the power of digitalization and big data to deliver the benefits of financial literacy to those who need it most.

Previous studies have highlighted a considerable discrepancy between financial knowledge and the decision-making abilities of lay individuals and financial experts. The consequences of a lack of financial knowledge can be significant. For instance, individuals without adequate financial literacy often misunderstand concepts like compounding, which may lead to misconceptions about key future outcomes like retirement security (Akerlof & Shiller, 2010). Other studies demonstrate that financial knowledge can affect investment risk perceptions (Diacon, 2004; Wang et al., 2011). In recent years, studies have highlighted a substantial deficiency in financial literacy in populations worldwide and called for better assessments and interventions (Aren & Dinç Aydemir, 2014; Atkinson & Messy, 2012). What is more remarkable is that the benefits of financial literacy training tend to be disproportionately distributed across demographic and socio-economic groups. For instance, studies conducted in the United States have consistently shown that low-income and less-literate households face barriers to accessing financial literacy programs programs even though these are the groups most likely to benefit from them (Braunstein & Welch, 2002; Dvorak & Hanley, 2010; Jacob et al., 2000). Hence, developing policy and training interventions that can lead to equitable growth in literacy is of key importance worldwide.

Many developed economies have, over the years, established commissions and constituted working groups to help assess the level of financial

[2] https://www.straitstimes.com/singapore/household-income-from-work-for-poor-families-fell-69-last-year-due-to-covid-19-study-by?close=true.

[3] https://www.straitstimes.com/opinion/virus-impact-to-still-weigh-on-economy-0.

literacy in their countries. National agencies like the Commission for Financial Capability[4] in New Zealand, Financial Capability[5] in the United Kingdom, Financial Consumer Agency of Canada (FCAC)[6] in Canada, Financial Literacy and Education Commission (FLEC)[7] in the United States, and Financial Capability[8] in Australia have been instrumental in developing and disseminating valuable information and tools to promote financial literacy. In Singapore, MoneySense was launched in 2003 as a national financial education program to help Singaporeans manage their finances and make sound financial decisions. MoneySense does so by providing free, practical, and unbiased financial education in a guided framework.[9] In 2012, MoneySense collaborated with Singapore Polytechnic to launch the Institute for Financial Literacy (IFL). IFL currently organizes programs to promote financial literacy and capability among Singaporeans through structured courses, seminars, workshops, and one-on-one literacy clinics. The aim of these programs is to provide free and unbiased training to help participants develop capabilities and competencies in key areas of financial planning such as money management, insurance, loans, credit, and retirement planning. IFL is increasingly looking to incorporate digital touchpoints and data-informed training strategies in how it offers its courses in Singapore.

In this chapter, we briefly review the current literature on financial literacy and offer prescriptions for how a multi-disciplinary mindset, coupled with a focus on data-backed insights such as rigorous assessment frameworks, artificial intelligence (AI), and machine learning, could offer a promising way forward for the field. We focus on three key themes: (i) definition and key concepts, (ii) data-backed insights from empirical evaluations of IFL's programs, and (iii) a data-informed and technology-enabled framework to enhance financial literacy in the digital era. In reviewing recent advancements in this field, we draw on our own field experiences as well as empirical evidence gathered as part of a study on

4 https://cffc.govt.nz/.

5 https://www.fincap.org.uk/.

6 https://www.canada.ca/en/financial-consumer-agency.html.

7 https://home.treasury.gov/policy-issues/consumer-policy/financial-literacy-and-education-commission.

8 https://www.financialcapability.gov.au/.

9 https://www.moneysense.gov.sg/about-us/moneysense-framework?sc_lang=en.

the implementation of a national-level financial education program in Singapore. Our early insights reveal the growing importance of financial confidence as a key driver for behavioral intention among learners and also point to population-level heterogeneities in financial literacy improvements.

2 Related Work

2.1 *Defining Financial Literacy*

The U.S. Financial Literacy and Education Commission defines financial literacy as the ability to use knowledge and skills to manage financial resources effectively toward financial well-being (Huston, 2010). In more recent work, Lusardi and Mitchell (2014) define financial literacy in terms of *"individuals' ability to process financial information and make informed decisions about financial planning, wealth accumulation, pensions, and debt."* While there is no clear consensus on the scope of these definitions (Hung et al., 2011; Huston, 2010; Remund, 2010), it is evident that most definitions of literacy focus not only on knowledge but also on related aspects such as capability and intention of decision-making. It is therefore not surprising that, over time, the scope of financial literacy has expanded to also include associated notions of financial skill, aptitude, competence, financial capability, and behavior (Holzmann, 2010; Remund, 2010; Xu & Zia, 2012).

2.2 *Associated Factors*

High financial literacy is known to be associated with making prudent money-related decisions. On the other hand, low financial literacy is linked to higher consumer debt, a lower saving rate, and bankruptcy (Fox & Bartholomae, 2008). In their study, Bayer et al. (2009) highlight that attending retirement seminars leads to greater participation in and contributions to savings plans, and that this improvement is more pronounced for employees who are not compensated highly. Other studies show that retirement planners are more adept at wealth accumulation when they have high levels of financial literacy (Lusardi & Mitchell, 2007). Similarly, financial literacy is found to have positive associations not only with retirement planning (Bernheim, 1998; Lusardi & Mitchelli, 2007) but with mortgage planning (Moore, 2003) and debt management

(Calvet et al., 2009; Lusardi et al., 2011). Furthermore, financial literacy training can be particularly beneficial for resource-constrained communities. Prior work done in the United States highlights that low-education and low-income individuals and families are often the least financially literate (Dvorak & Hanley, 2010) and that there is a dearth of public and private programs available to them (Braunstein & Welch, 2002; Jacob et al., 2000). In their report, Jacob and colleagues (2000) argue that with the rapidly changing financial landscape and the increased penetration of digital finance services, low-income families in the United States lack basic financial skills, and are increasingly vulnerable to economic shocks such as job loss and medical emergencies (Jacob et al., 2000). However, while digital training combined with financial literacy programs can be beneficial, the results can often be mixed (Jacob et al., 2000). For instance, results from a field study and survey conducted in India and Indonesia show that a financial education program boosts demand for opening bank accounts, but only for respondents who had low levels of education or financial literacy (Cole et al., 2011).

Financial investing is another area of key importance. Studies in this area show that higher financial literacy is linked to greater participation in the stock market (van Rooij et al., 2011) and that financially sophisticated investors make fewer mistakes. For instance, more financially literate investors were more likely not to under-diversify their portfolios, e.g., (Guiso & Jappelli, 2008; van Rooij et al., 2011). Research also reveals that higher financial literacy is linked to portfolio diversification, stronger returns (Abreu & Mendes, 2010; Lusardi et al., 2017), and a higher likelihood of buying low-cost alternatives to actively managed funds (Müller & Weber, 2010).

3 *Learning About Learners*: Analyzing Participant Behavior in a Large-Scale Financial Literacy Program

As outlined in the previous section, in partnerships with various organizations in Singapore, IFL has been conducting financial literacy modules in the form of workshops and talks. At the time of writing this chapter, IFL has conducted a total of 3361 workshops and 2524 talks in-person, as well as 363 talks and 365 workshops online. Before beginning each workshop, participants are asked to complete a form that records their demographic information and their current state of financial knowledge and confidence. Upon completing the workshop, participants are asked to complete another form that records their overall appraisal of the module and trainer, their interests in other financial topics, their post-workshop or "new" state of financial knowledge and financial confidence, and lastly, their intention to act on the specific topic of the module. As such, changes in financial knowledge and financial confidence can be assessed by comparing the corresponding responses before and after the workshop.[10] The availability of this large and fine-grained dataset on learner background, preferences, and experiences with IFL's modules offers an opportunity for us to implement learning analytics. Next, we present some early aggregated insights from analyzing this anonymized feedback data comprising 70,000 observations spanning May 2016 to November 2018 collated by IFL. These insights can help us better understand the effectiveness of financial literacy workshops for a large sample of participants and offer prescriptive guidance for the design of personalized digital interventions in future. Some of the key findings from this analysis are likely to apply beyond Singapore and to similar financial literacy training programs and learners in other developed economies worldwide. Table 1 presents the number of participants in each of the modules across workshops and talks.

Figure 1 illustrates the demographic distribution of participants across the IFL workshops and talks. Based on the dataset, we find that, when compared to the population of Singapore, our data over-represents males, individuals with post-secondary level education, and individuals in the age group of 21–30 years. The dataset under-represents females, individuals with secondary level education and below, and individuals who are below 21 years or above 30 years of age.

Table 2 shows the pairwise correlation for our dataset's key variables, including the demographic variables listed above, module-related variables (i.e., content evaluation, trainer evaluation, and interest topics), and outcomes variables (i.e., financial knowledge, financial confidence, and intention).

[10] We note that these feedback exercises are strictly optional, and no personal identifiers are collected as a general policy. However, participants do have the option of entering their email addresses if they wish to subscribe to a monthly newsletter.

Table 1 Summary of participation in the Institute for Financial Literacy (IFL) workshops and talks

*Module topic/title**	*No. of Participants*	
	Talk	*Workshop*
Debt management	**4,544**	**1,174**
Understanding loans and credit	4,544	1,174
Financial planning	**4,500**	**15,728**
Implementing your financial plan	1,533	–
Importance of estate planning	904	–
Measuring your financial fitness	1,694	–
Steps in estate planning	369	–
Basics of money management	–	27
Building your nest egg	–	3,939
Buying a home within your means	–	510
Financial planning begins now	–	3,361
Financial planning for home team NSFs	–	5,597
Introduction to estate planning	–	1,957
Starting a family	–	337
Health insurance	**478**	**624**
Types of health insurance	478	–
Understanding basic health insurance	–	624
Insurance planning	**3,851**	**5,091**
Assessing your insurance needs	251	–
CPF insurance schemes	3,188	–
Types of life insurance	412	–
Do I need every type of insurance	–	4,746
Understanding life insurance	–	345
Investment planning	**3,306**	**5,371**
Considerations when investing	1,792	–
Major financial products	1,165	–
Understanding bonds	349	302
Fundamentals of share investing	–	612
Introduction to personal investing	–	4,239
Understanding exchange traded funds (ETFs)	–	68
Understanding real estate investment trusts (REITS)	–	150
Money management	**4,291**	**4,743**
Baby and child support schemes	318	–
Beware of scams	1,146	–
Budgeting	1,106	–

(continued)

Table 1 (continued)

*Module topic/title**	*No. of Participants*	
	Talk	***Workshop***
Stretching your dollar	1,554	–
When the deal turns sour	167	–
Is the deal too good to be true	–	664
Making sense of your money	–	3,717
Money management for youth	–	362
Retirement planning	**4,558**	**3,289**
Assessing your retirement income needs	720	–
Options to build your retirement income	1,166	–
Managing your CPF money for retirement	2,672	2,639
Enrich your golden years	–	650
Customized topics	**7,580**	**1,608**
Sub-total	**33,108**	**37,628**
Grand-total	**70,736**	

*Some module titles are illustrative of the general topic. The exact phrasing of title might differ depending on workshop context.

3.1 Financial Literacy Training Boosts Financial Knowledge, Confidence, and Intention

Based on a comparison of post-versus pre-session feedback scores, we analyzed the change in participants' self-reported financial knowledge and financial confidence following their participation in an IFL workshop.[11] Figure 2 illustrates the distribution of knowledge and confidence scores for pre-and post-workshop sessions, while Fig. 3 illustrates the distribution of knowledge and confidence increase/decrease from attending the workshops.

Figures 2 and 3 show that most participants report a substantial improvement in knowledge and confidence after attending an IFL workshop. The observed peaks at scores 5, 10, 15, and 20 are likely due to participants answering either 1, 2, or 3, etc., to each of the five items in the question. However, the distribution of pre-and post-workshop knowledge scores shows a fairly uniform distribution. Figure 3 also reveals that 23.8% (6,891) and 22.3% (6,218) of participants, respectively, report no

[11] For this analysis, we use observations exclusively from IFL workshops since pre-module financial knowledge and confidence indicators were not collected for IFL talks.

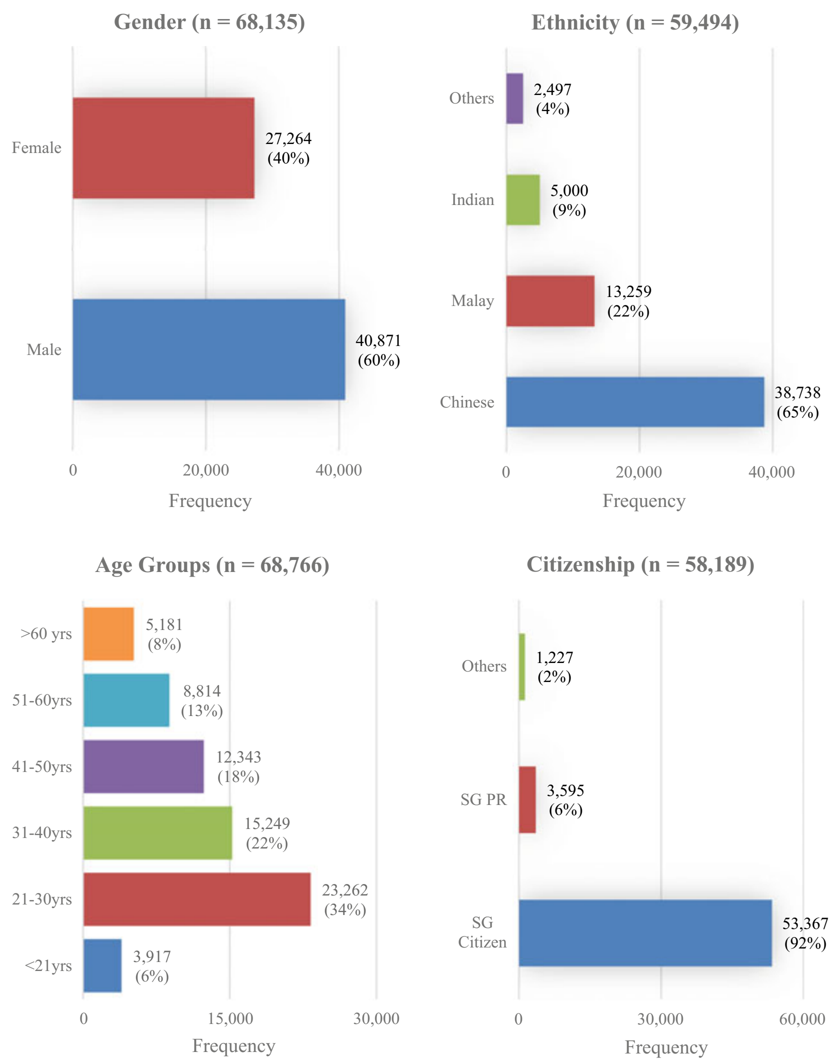


Fig. 1 Demographic distribution of participants

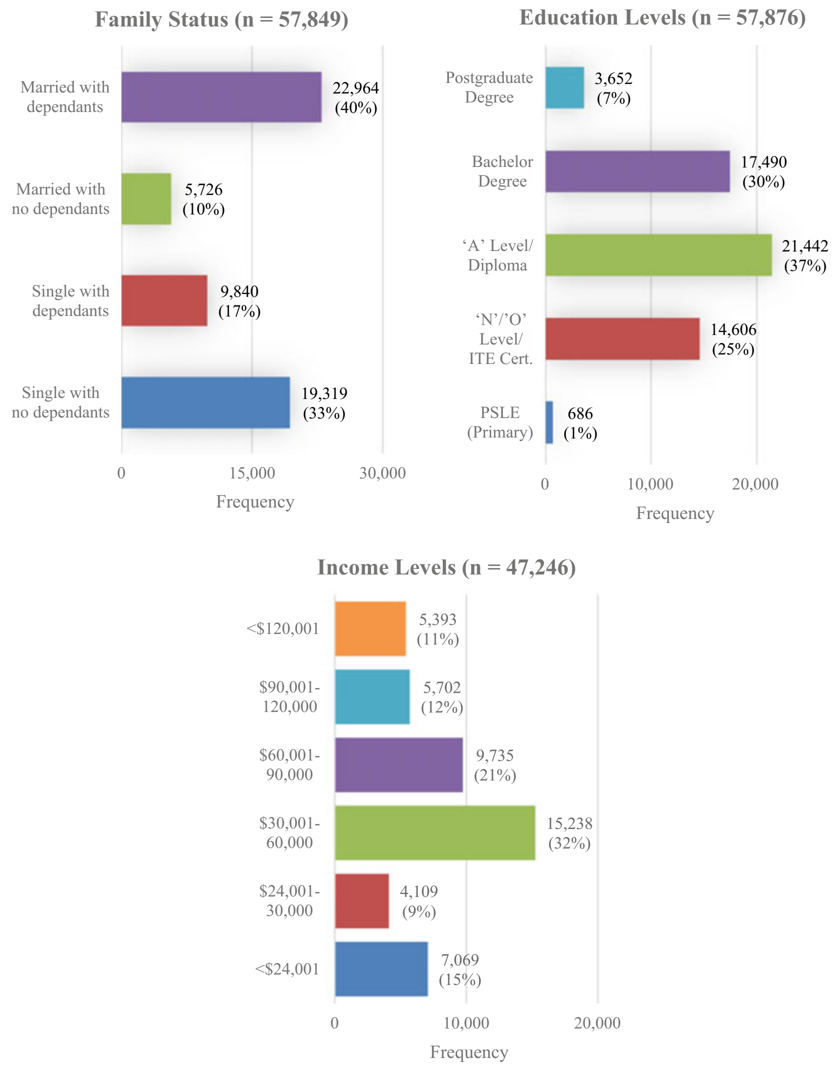


Fig. 1 (continued)

Table 2 Correlation matrix for workshops (by pairwise deletion)

	n	M	SD	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
1. Age group	36,688	3.04	1.27	–									
2. Education level	34,252	3.03	0.85	−0.02	–								
3. Income level	28,880	3.32	1.50	0.35	0.38	–							
4. Content evaluation—objective	33,558	4.18	0.69	0.02	0.03	0.03	–						
5. Content evaluation—useful	33,545	4.17	0.70	0.02	0.05	0.03	0.82	–					
6. Content evaluation—example	33,517	4.18	0.68	0.01	0.03	0.03	0.80	0.81	–				
7. Content evaluation—concept	33,525	4.16	0.69	0.02	0.03	0.03	0.79	0.80	0.82	–			
8. Content evaluation average	33,383	4.18	0.64	0.02	0.04	0.03	0.92	0.93	0.93	0.92	–		
9. Trainer evaluation—knowledge	33,534	4.38	0.70	0.01	0.03	0.02	0.72	0.71	0.73	0.73	0.78	–	
10. Trainer evaluation—communicate	33,528	4.3	0.70	0.03	0.03	0.03	0.73	0.72	0.76	0.75	0.80	0.84	–
11. Trainer evaluation—topic	33,523	4.25	0.71	0.03	0.03	0.03	0.74	0.75	0.77	0.78	0.82	0.78	0.84
12. Trainer evaluation—example	33,485	4.26	0.70	0.03	0.02	0.03	0.73	0.72	0.77	0.75	0.81	0.79	0.83
13. Trainer evaluation average	33,363	4.3	0.65	0.03	0.03	0.03	0.78	0.78	0.81	0.81	0.86	0.92	0.94
14. Pre-workshop knowledge	33,341	2.11	1.48	−0.03	0.14	0.05	0.06	0.05	0.06	0.06	0.06	0.05	0.06
15. Post-workshop knowledge	30,186	3.41	1.37	−0.07	0.23	0.10	0.14	0.14	0.13	0.13	0.14	0.15	0.14
16. Post-pre workshop knowledge difference	28,936	1.32	1.60	−0.03	0.07	0.04	0.05	0.06	0.05	0.05	0.06	0.07	0.06
17. Pre-workshop confidence	32,851	11.12	3.82	−0.11	−0.08	−0.08	0.09	0.07	0.08	0.09	0.09	0.06	0.08
18. Post-workshop confidence	29,466	14.27	3.02	−0.12	0.00	−0.03	0.27	0.25	0.27	0.27	0.29	0.22	0.25
19. Post–pre workshop confidence difference	27,907	3.15	3.36	0.02	0.09	0.06	0.14	0.15	0.15	0.14	0.16	0.14	0.14
20. Post-workshop intention	29,588	19.18	3.15	−0.07	0.07	0.05	0.21	0.21	0.21	0.21	0.23	0.17	0.19
21. Indicated interest topic	37,628	1.98	1.89	−0.10	0.07	−0.03	0.06	0.07	0.06	0.07	0.07	0.06	0.06

(continued)

Table 2 (continued)

	n	M	SD	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*
1. Age group	36,688	3.04	1.27											
2. Education level	34,252	3.03	0.85											
3. Income level	28,880	3.32	1.50											
4. Content evaluation—objective	33,558	4.18	0.69											
5. Content evaluation—useful	33,545	4.17	0.70											
6. Content evaluation—example	33,517	4.18	0.68											
7. Content evaluation—concept	33,525	4.16	0.69											
8. Content evaluation average	33,383	4.18	0.64											
9. Trainer evaluation—knowledge	33,534	4.38	0.70											
10. Trainer evaluation—communicate	33,528	4.3	0.70											
11. Trainer evaluation—topic	33,523	4.25	0.71	–										
12. Trainer evaluation—example	33,485	4.26	0.70	0.84	–									
13. Trainer evaluation average	33,363	4.3	0.65	0.93	0.93	–								
14. Pre-workshop knowledge	33,341	2.11	1.48	0.06	0.05	0.06	–							
15. Post-workshop knowledge	30,186	3.41	1.37	0.13	0.13	0.15	0.37	–						
16. Post–pre workshop knowledge difference	28,936	1.32	1.60	0.05	0.06	0.06	−0.61	0.51	–					
17. Pre-workshop confidence	32,851	11.12	3.82	0.08	0.08	0.08	0.43	0.08	−0.32	–				
18. Post-workshop confidence	29,466	14.27	3.02	0.26	0.25	0.26	0.26	0.24	−0.05	0.54	–			
19. Post–pre workshop confidence difference	27,907	3.15	3.36	0.15	0.14	0.15	−0.25	0.12	0.34	−0.65	0.29	–		
20. Post-workshop intention	29,588	19.18	3.15	0.21	0.19	0.21	0.24	0.26	−0.01	0.37	0.54	0.06	–	
21. Indicated interest topic	37,628	1.98	1.89	0.07	0.06	0.06	0.03	0.14	0.08	−0.01	0.03	0.04	0.07	–

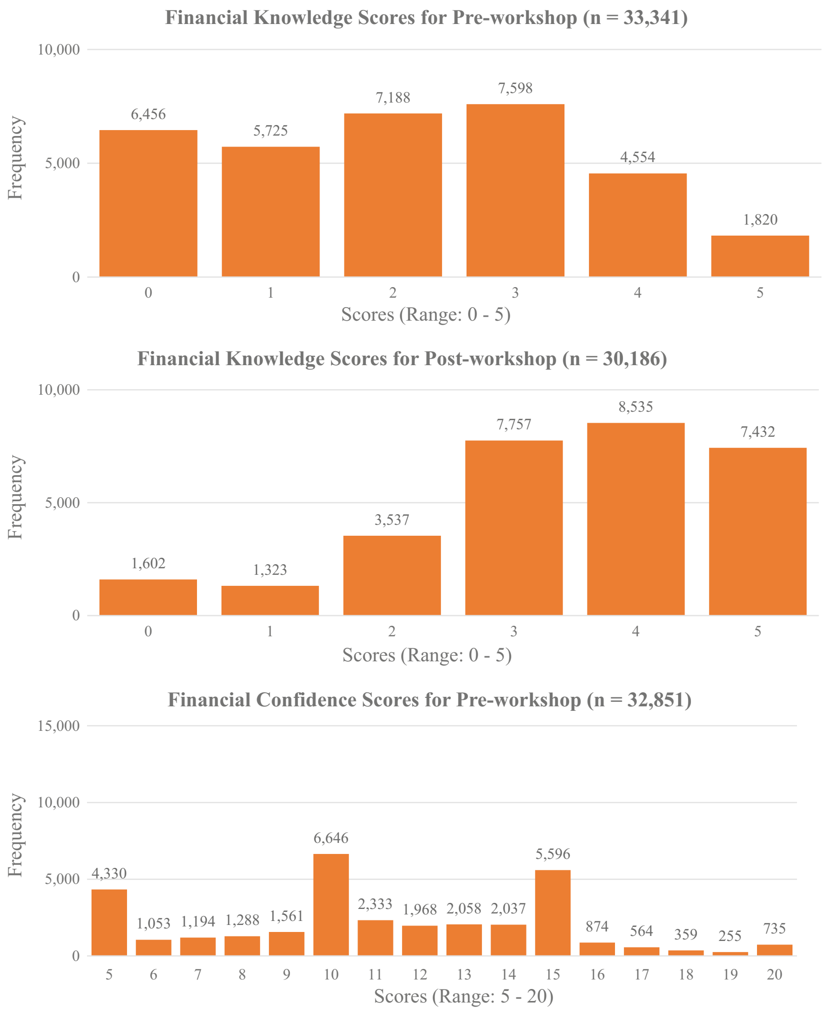


Fig. 2 Distribution of financial knowledge and financial confidence scores pre- and post-workshop

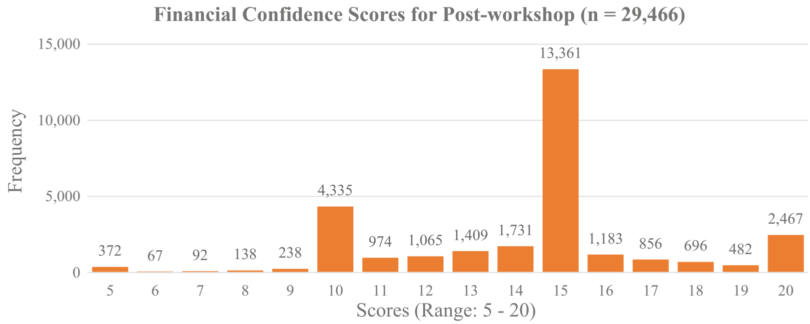
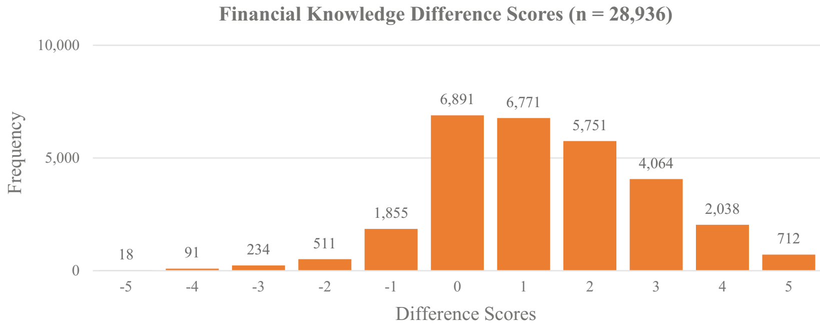


Fig. 2 (continued)



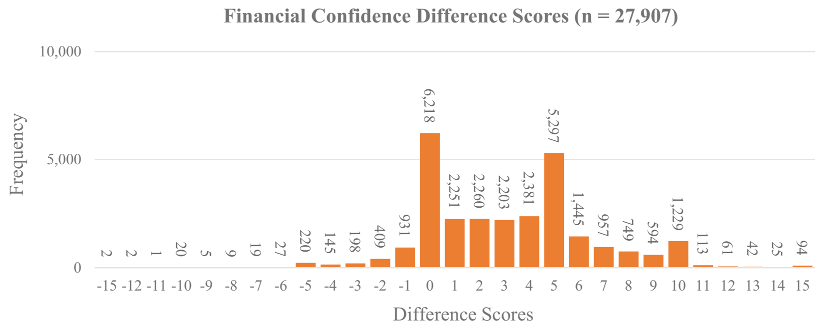


Fig. 3 Distribution of financial knowledge and financial confidence difference (increase/decrease) scores post- and pre-workshop

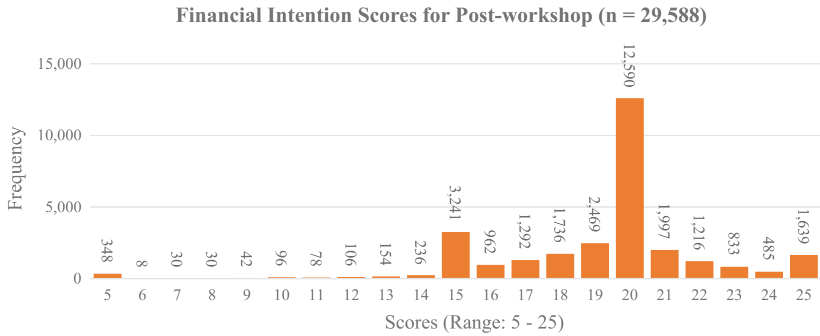


Fig. 4 Distribution of post-workshop financial intention

change in knowledge and confidence scores after attending a workshop. This result is a significant percentage and suggests that the potential exists for better targeting of participants to improve their learning outcomes. Interestingly, we also note that 7.2% (2,079) and 7.1% (1,988) of participants reported a reduction in financial knowledge and confidence after attending a workshop. It would be important to understand the reasons for this and what can be done to engage this group of learners effectively. Figure 4 presents the distribution of financial intention scores as reported by the participants upon completing the workshop. As our findings make evident, most participants report a moderate to high intention to act on their financial goals upon completing the financial workshop.

3.2 Heterogeneities in Impact of Financial Literacy Trainings

As described in the previous section, attending an IFL workshop is beneficial to most participants in terms of improving financial knowledge and confidence, as well as eliciting a positive intention to act. However, we also note that these beneficial effects vary across participant groups and across module topics.

Variance across age groups: From Table 2, we see that age has a significantly negative correlation with both pre-workshop confidence ($r = -0.11$, $p < 0.001$) as well as post-workshop confidence ($r = -0.12$, $p < 0.001$). On further inspection, we find that age has a curvilinear relationship with confidence, which implies that confidence decreased with age from <21 years to 41–50 years but increased from 51 to 60 years onward. This result is visually illustrated in Fig. 5. Interestingly, we can also see that age negatively correlates with total indicated interest ($r = -0.14$, $p < 0.001$), which measures the total number of topics in which the participant has expressed interest. We also find that age has an inverse curvilinear relationship with total indicated interest, where the mean total indicated interest increases with age from <21 years to 21–30 years but subsequently decreases. This result is illustrated in Fig. 5. As individuals age, the number of relevant life events (e.g., marriage, childbirth) becomes fewer, and this is likely reflected in a shrinking set of interest topics over time.

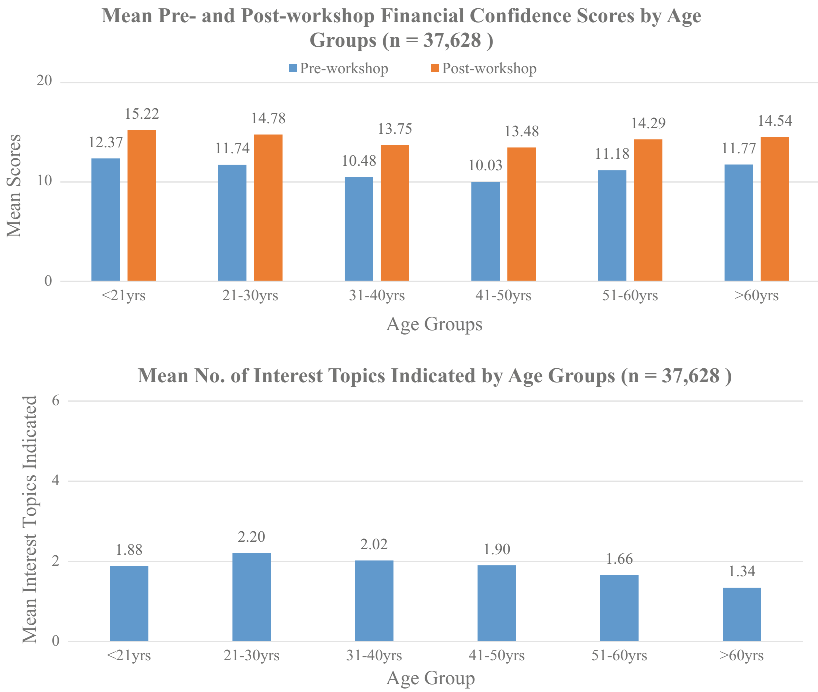


Fig. 5 Mean pre- and post-workshop financial confidence scores and mean number of interest topics indicated, across age groups

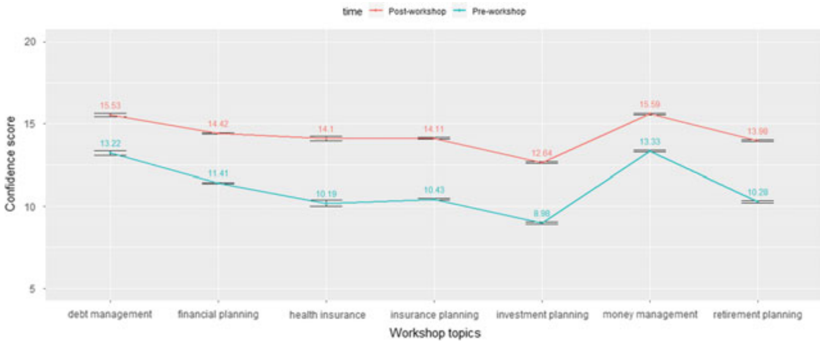


Fig. 6 Mean financial confidence scores pre- and post-workshop across topics

Variance across topics: We note that there also exist significant heterogeneities in confidence gain across IFL's workshops based on the topic.[12] We show the mean confidence for pre-and post-workshop across topics in Fig. 6, revealing significant variance. A mixed-design Analysis of variance (ANOVA) analysis was adopted to quantify the effect sizes and verify if these changes were statistically significant. Our results show that the main effect for time is significant, $F(1, 27900) = 8417.11$, $p < .001$, with a fairly large partial effect size of 0.23. The difference in mean confidence between post- and pre-workshop sessions is 3.15. The estimate for interaction between time and course topic is also significant, $F(6, 27900) = 84.32$, $p < .001$, but with a small partial effect size of 0.02. In addition, a post hoc pairwise comparison between post- and pre-workshop sessions by topics (Table 3) reveals that the reported confidence is higher post-workshop than pre-workshop across topics. Interestingly, the largest and smallest confidence differences are observed for the topics of health insurance and money management, respectively.

[12] The specific survey items that measured pre- and post-workshop confidence were contextualized to the specific topic of the workshops. The findings of a comparison of overall confidence scores across different workshops might therefore be unreliable.

Table 3 Post-hoc pairwise comparison of financial confidence scores post- and pre-workshop by module topics

Module Topics	*Mean Difference*	df	t
Debt management	2.31	830	19.90***
Financial planning	3.01	12,252	102.00***
Health insurance	3.91	331	21.40***
Insurance planning	3.68	4179	67.70***
Investment planning	3.66	3993	67.10***
Money management	2.26	3830	45.80***
Retirement planning	3.70	2485	52.70***

Method: Bonferroni correction; *Note* $*p < .05$, $**p < .01$, $***p < .001$

3.3 Factors Associated with Financial Intention

In this section, we look at the association between key workshop-outcome variables (i.e., financial knowledge and financial confidence) and the self-reported intention of participants to act. Table 4 shows that intention has the highest positive correlation with post-workshop confidence ($r = .54$, $p < .001$) while having a lower but positive correlation with post-workshop knowledge ($r = .26$, $p < .001$). We regress intention on post-workshop knowledge and post-workshop confidence, and find the overall model to be significant ($F(2, 28651) = 6505$, $p < .001$) with a $R^2 = .31$. Furthermore, the results show that financial knowledge ($b = 0.35$, $t(28299) = 28.47$, $p < .001$) and financial confidence ($b = 0.53$, $t(28299) = 100.66$, $p < .001$) significantly affect intention, after controlling for variables relating to workshop evaluation, demographic, and socio-economic indicators (i.e., gender, age, education level, income level, and ethnicity), and workshop topics. Notably, post-workshop confidence, in particular, has a much larger proportional effect on the variance in intention, than other independent variables. This result is consistent with findings from other countries highlighting the importance of financial confidence as a key determinant of financial behavior, at all levels of financial knowledge (Asaad, 2015). In conjunction with Singapore

Polytechnic's pedagogical expertise, this key finding led IFL to develop a Confidence Framework. This framework guides IFL's review of its modules and training delivery to assess how the financial confidence of participants could be improved. The framework advances four key elements through which financial confidence can be boosted: defining purpose, building a community, giving choices, and providing space for simulated practice. As a result, various aspects of the financial literacy training offered by IFL are currently in the process of being improved. For example, electronic self-directed learning versions of IFL's modules are currently being produced to give participants the choice of pursuing financial literacy at their own pace alongside existing training modalities such as online workshops. Also, training content delivery has been revised to provide participants attending online/in-person training with clear learning objectives.

3.4 *Generating Profiles of IFL's Learners Using Cluster Analysis*

To better understand the profiles of participants who attended IFL's programs (i.e., workshops and talks), we conducted a simple cluster analysis to classify participants based on their demographic and socio-economic indicators, as well as the module topics they chose to attend. Cluster analysis techniques have been widely used in research to summarize big data to create meaningful insights by categorizing the data into a set of groups or clusters based on similarity (or dissimilarity) between the entities (Fahad et al., 2014; Greaves, 2019). For our analysis, we use a k-medoids clustering approach using a Gower distancing metric to generate the clusters, and t-distributed stochastic neighbor embeddingedding (t-SNE) to generate a low-dimensional representation of the clusters for better interpretation. The variables in our model include[13] *module topics, gender, ethnicity, age groups, marital status, education level, and income level*. Given the computational limitations of a large dataset ($n = 45{,}775$, after case-wise deletion), we randomly sample 30,000 observations from the dataset for this analysis. As a final consideration, we determine the optimal number of clusters to generate. We use silhouette width as a quantitative indicator of clustering quality and obtain the highest silhouette width score for the 3-cluster and 7-cluster solutions. The t-SNE

[13] As citizenship did not have much variation, it was omitted from the cluster analysis.

Table 4 Regressing financial intention on key factors after controlling for module evaluation, demographics, and topics

Variables	*Estimate*	*S.E*	*sr^2*	*t*	*df*	*F*	*adj R^2*
Model 1					(2, 28,651)	6505***	0.31
Post-module knowledge	0.35	0.012	0.03	28.47***			
Post-module confidence	0.53	0.005	0.26	100.66***			
Model 2					(4, 28,299)	3266***	0.32
Post-module knowledge	0.34	0.013	0.03	27.34***			
Post-module confidence	0.51	0.006	0.23	93.14***			
Content evaluation average	0.30	0.048	0.001	6.25***			
Trainer evaluation average	0.04	0.047	0.00002	0.79			
Model 3					(11, 22,956)	997.40***	0.32
Post-module knowledge	0.34	0.01	0.02	23.09***			
Post-module confidence	0.51	0.01	0.23	83.39***			
Content evaluation average	0.31	0.05	0.002	5.82***			
Trainer evaluation average	0.03	0.05	0.00002	0.6			
Age	−0.02	0.02	0.00005	−1.07			
Education level	0.09	0.02	0.0006	3.80***			
Income level	0.08	0.01	0.002	6.03***			
Gender—male (vs. female)	0.18	0.04	0.0009	4.51***			
Ethnicity—Malay (vs. Chinese)	0.05	0.04	0.00007	1.29			
Ethnicity—Indian (vs. Chinese)	−0.01	0.06	0.000001	−0.14			
Ethnicity—Others (vs. Chinese)	0.13	0.08	0.0001	1.64			

(continued)

Table 4 (continued)

Variables	*Estimate*	*S.E*	sr^2	*t*	*df*	*F*	*adj* R^2
Model 4					(17, 22,950)	670.80***	0.33
Post-module knowledge	0.33	0.01	0.02	22.28***			
Post-module confidence	0.50	0.01	0.22	79.99***			
Content evaluation average	0.33	0.05	0.002	6.24***			
Trainer evaluation average	0.01	0.05	0.000003	0.28			
Age	0.04	0.02	0.0003	2.53*			
Education level	0.09	0.02	0.0007	3.87***			
Income level	0.10	0.01	0.002	7.00***			
Gender—male (vs. female)	0.15	0.04	0.0006	3.72***			
Ethnicity—Malay (vs. Chinese)	0.08	0.04	0.0002	1.89			
Ethnicity—Indian (vs. Chinese)	0.00	0.06	0.00	0.002			
Ethnicity—Others (vs. Chinese)	0.16	0.08	0.0002	1.92			
Topic—debt management (vs. financial planning)	−1.10	0.10	0.005	−10.69***			
Topic—health insurance (vs. financial planning)	0.11	0.15	0.00002	0.71			
Topic—insurance planning (vs. financial planning)	0.20	0.05	0.0007	3.92***			
Topic—investment planning (vs. financial planning)	−0.51	0.05	0.004	−9.46***			
Topic—money management (vs. financial planning)	0.33	0.06	0.002	6.04***			

(continued)

Table 4 (continued)

Variables	*Estimate*	*S.E*	*sr*^{2}	*t*	*df*	*F*	*adj* R^2
Topic—retirement planning (vs. financial planning)	−0.19	0.07	0.0003	−2.81***			

Note *p < 0.05, **p < 0.01, ***p < 0.001

plot in Fig. 7 illustrates the cluster distribution for the 3-cluster solution. The data points across clusters show significant overlap, suggesting that the clusters are not as cleanly defined. Nevertheless, we still observe some qualitative distinctions among the clusters as certain demographic, socio-economic, or course-attended factors are over-represented in certain clusters, as illustrated in Table 5.

Specifically, cluster 2 appears to represent a profile of young working adults who typically attend basic financial courses such as debt management, financial planning, insurance planning, and money management. In contrast, clusters 1 and 3 represent a profile of older working adults who typically attend courses that serve specialized needs such as health insurance, investment planning, and retirement planning. However, these

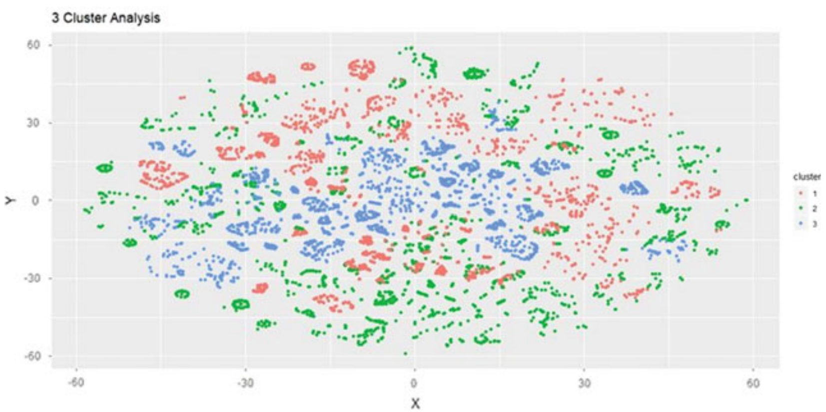


Fig. 7 3-cluster solution visualization

Table 5 Profile summary of 3-cluster solution

	Cluster #1	*Cluster #2*	*Cluster #3*
n (proportion)	9,516 (31.72%)	12,136 (40.45%)	8,348 (27.83%)
Attended module topics	Financial Planning Health Insurance Investment Planning Retirement Planning	Debt Management Financial Planning Insurance Planning Money Management	Financial Planning Health Insurance Insurance Planning Investment Planning Retirement Planning
Gender	All Male	Mostly Male	All Female
Ethnicity	Chinese Indian Malay	Chinese Malay Others	Chinese
Age groups	31–40yrs 41–50yrs 51–60yrs >60yrs	<21yrs 21–30yrs 31–40yrs	21–30yrs 31–40yrs 41–50yrs 51–60yrs >60yrs
Family status	Married no dependent Married with dependent	Single no dependent Single with dependent	Married no dependent Married with dependent Single with dependent
Education level	PSLE (Primary) 'N' Level/'O' Level/ITE 'A' Level/Diploma Bachelor Degree Postgraduate Degree	'N' Level/'O' Level/ITE 'A' Level/Diploma Bachelor Degree	'A' Level/Diploma Bachelor Degree Postgraduate Degree
Income levels	\$30,001–60,000 \$60,001–90,000 \$90,001–120,000 >\$120,001	<\$24,001 \$24,001–30,000 \$30,001 & \$60,000	\$30,001–60,000 \$60,001–90,000 \$90,001–120,000 >\$120,001

clusters differ in that individuals from cluster 1 are all male while individuals in 3 are all female. This clustering exercise helps reveal underlying similarities and differences in learner characteristics across workshops and hints at the possibility of designing data-informed personalization strategies that can better target the learning needs of these participants. Furthermore, this exercise highlights the relevance of using big data analytic methods to analyze fine-grained data on learner profiles and behavior. This exercise is consistent with studies in learning analytics that have used both supervised and unsupervised learning approaches to uncover insights into how students learn.

4 Recommendations for a Data-Informed Financial Literacy Program

In the previous sections, we have illustrated key concepts and factors related to financial literacy. We have also presented insights from an analysis of participant feedback data collected from IFL's past workshops. Drawing on these learnings, we present four recommendations for developing a data-informed strategy for enhancing financial literacy.

4.1 *Expanding Financial Literacy Touchpoints on Mobile and Web*

Ever since the beginning of the COVID-19 pandemic in early 2020, governments worldwide have introduced restrictions on public gathering and social distancing protocols. These measures have, in turn, pushed organizations and businesses to quickly adopt digitalization strategies to engage with employees and customers (Almeida et al., 2020; Faraj et al., 2021). Recent reports suggest that the average share of online global customer interactions increased by over 20% between December 2019 and July 2020 alone.[14] It is evident that the next few years will continue to witness accelerated digitalization across sectors and services. Singapore has also increased financial literacy engagements through its existing digitalization initiatives in the public and private sectors. For example, MyMoneySense is a financial planning digital service developed by the Ministry of Manpower in collaboration with the Government

[14] https://www.mckinsey.com/business-functions/strategy-and-corporate-finance/our-insights/how-covid-19-has-pushed-companies-over-the-technology-tipping-point-and-transformed-business-forever.

Technology Agency of Singapore, which leverages the Singapore Financial Data Exchange (SGFinDex) launched by the Monetary Authority of Singapore and the Smart Nation and Digital Government Group. MyMoneySense aims to provide unbiased, personalized, and actionable financial guidance for Singaporeans at all stages of life. IFL has also added to its Confidence Framework by building an online community of learners that facilitates the transfer of tacit knowledge. This community has materialized through IFL's launch of a moderated community platform called FinLearn Community. Participants in IFL training are invited to register and join the community in order to be able to continue discussions on financial literacy topics and related topics. These discussions allow IFL to understand better learner perceptions and the topics in which there is emerging interest. Alongside FinLearn Community, IFL is also currently in the process of digitalizing its financial literacy training content using an Electronic Learning Management System (ELMS), also called FinLearn. When completed, FinLearn is intended to be a one-stop resource for participants to manage their training and self-directed learning goals. It would also help automate all knowledge, confidence, and intention assessments. Furthermore, FinLearn would also enable more accurate and reliable measurements of participant actions in response to the training and thus benefit our learning analytics initiatives described earlier.

4.2 Development of a Psychologically Enhanced Learner Profile for Improved Personalization

Earlier, we described how a big data-based clustering approach could be used to uncover certain demographic, socio-economic, and behavioral profiles of IFL learners. An important extension of this line of work would be to augment the profiling system by incorporating data on the relevant psychological profiles of individuals. Based on a brief review of the literature on the interplay between psychological constructs and financial decision-making, the following three constructs can be considered illustrative examples.

- ***Risk-taking***: captures differences in the willingness of individuals to take risks (Glass et al., 1965; Zuckerman, 2000). Within the financial sector, assessing risk profiles as a combination of risk-taking and risk-averse capacity and behavior is a common practice. This assessment is particularly true in the field of investments and wealth accumulation

(Klement, 2015). According to a recent study, men with a higher risk tolerance are more likely to associate positively with standard and sophisticated investment decisions, while women only associate positively with standard investment decisions (Bannier & Neubert, 2016). In another study, Cavezzali et al. (2012) show that financial literacy, in turn, positively affects individuals' risk-taking decisions but only partially explains their choice of diversification strategies. The study also makes prescriptive suggestions on how to improve financial literacy for potential investors. Similar findings have also been reported in other countries (Bajo et al., 2015).

- ***Time perspective***: captures the inherent tendency of individuals to focus on and dwell in the past versus living in the present or planning for the future (Collingwood, 2016). Researchers have identified multiple types of time orientations that individuals may possess and display (P.G. Zimbardo & Boyd, 1999). Time perspectives have been known to influence various aspects of quality of life, including general well-being, ability to resolve conflicts, and professional success (Zhang et al., 2013; P. Zimbardo & Boyd, 2008). Recent studies have explored the interplay between financial literacy, time perspective, and other associated outcomes. For instance, in a study conducted on 18–35-year-olds, researchers show that the different time orientations can affect the ability to delay gratification, and thus affect financial well-being (Zsótér, 2018). It has also been argued that including a consideration of time perspective in financial literacy training is critical to improving overall financial health (P. Zimbardo et al., 2017).
- ***Need for cognition (NFC)***: measures an innate tendency/willingness to engage in activities that require elaborate thinking (e.g., solving a complex puzzle) (Cacioppo & Petty, 1982; Osberg, 1987). High NFC individuals prefer high elaborations, while low NFC individuals prefer heuristics and short-cuts. NFC can also mediate or moderate the effect of various phenomena on consumer behaviour. For instance, Lin et al. (2011) report that NFC moderates the effect of online product reviews on purchase intention.

Based on the above psychological constructs, we propose a possible framework for measuring psychological profiles and administering financial literacy interventions using a digital platform (see Fig. 8). Users

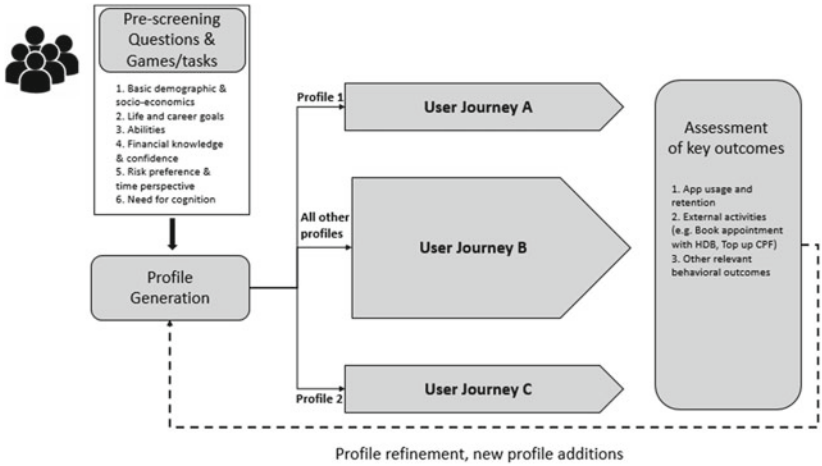


Fig. 8 User profiling and intervention analysis framework for financial literacy enhancement

logging on to the platform are required to complete an initial pre-screening questionnaire or an equivalent set of tasks/games, from which the platform will generate a basic profile (i.e., based on demographic and socio-economic indicators) as well as an enhanced profile (i.e., basic profile + psychological constructs) of each user. Based on this enhanced profile, the platform will create a personalized learning experience, routing individuals through specific user journeys. At intermediate points, as well as at the end of the journey, we can assess key learning and behavioral outcomes.

4.3 Continuous Evaluation of Financial Literacy Programs and Policies

Drawing on the digital intervention analysis framework presented in the previous section, it is evident that accurate assessments of the efficacy of financial literacy programs are critical to improving decision-making and policy guidance (Kaiser et al., 2022). Past initiatives and studies offer useful guidance on assessment scales (OECD International Network on Financial Education, 2011), assessment frameworks (Blue & Grootenboer, 2019), intervention types (Hathaway & Khatiwada, 2008), and cautionary notes (Gale & Levine, 2013). Specifically, sample surveys conducted at repeated intervals can help establish and refine population benchmarks on financial knowledge, confidence, capability, and intention. To better understand changing preferences and motivations, researchers should conduct structured or semi-structured interviews at an individual- or household level. In addition to the descriptive analyses detailed above, we also recommend greater use of prescriptive analytics that use causal inference models to infer the impact of financial literacy programs on financial intentions and behavior. Several frameworks of behavioral change have been proposed to assist researchers to be able to design effective strategies and techniques to guide changes in behavior. The Behavioral Change Wheel (Michie et al., 2016; "BCW"), illustrated in Table 6 below, is one such framework that follows a holistic three-stage and eight-step approach. In short, we believe that the combination of a (i) digital touch-point for continuous measurement of learner engagement, (ii) design and administration of digital interventions (e.g., app-based study alerts), (iii) big data storage and processing capabilities, and (iv) rigorous assessment techniques can offer a flexible and scalable way to enhance financial literacy, not just in Singapore, but globally.

Table 6 Designing behavior change interventions

Stage 1	Behavioral understanding	In this phase, researchers define the problem in behavioral terms, select a specific behavior as a potential target for change, and clearly specify how this behavior needs to change
Stage 2	Identifying interventions	In this phase, researchers identify suitable intervention functions (e.g. education, persuasion, incentives) and associated policy changes that can be used to facilitate the target behavior change
Stage 3	Content and implementation	In the final phase, researchers finalize the specific behavioral change techniques that need to be implemented and delivered

Source Adapted from Michie et al. (2016)

4.4 Increased Focus on Specific Population Groups

Financial literacy training and programs have been particularly beneficial for resource-constrained communities worldwide. Numerous studies have emphasized that low-income and low-literacy families tend to have low financial literacy levels and might be more likely to make poor financial decisions (Dvorak & Hanley, 2010). Ironically, studies show that households that are most in need of financial literacy interventions are also the ones that lack access to them (Braunstein & Welch, 2002; Jacob et al., 2000). In the absence of adequate financial literacy, these individuals and families remain at considerable risk in the face of financial shocks like job loss or economic downturns. We contend that financial literacy programs should first and foremost target members of society who are at the highest risk of financial distress and/or those who lack access to such programs owing to various factors such as time constraints, low-literacy skills, or digital exclusion. Moreover, while financial literacy programs need to be part of a universal literacy policy, the urgent need for and the effectiveness of such programs will likely vary substantially across specific groups and individuals in the population. This heterogeneity can be better understood through careful measurement and analysis of the available data on literacy outcomes.

5 Conclusion

Due to the rapid emergence of new financial products and services, as well as constantly evolving policies on savings and retirement, there is an increased and urgent need for citizens to be financially literate. In this chapter, we summarize a discussion of recent work in financial literacy from across the world, illustrate preliminary insights from a large sample of past financial literacy workshops conducted with participants in Singapore, and discuss recommendations for an effective and data-informed financial literacy enhancement strategy. We have also sought to emphasize the need to adopt a systematic approach to financial literacy enhancement that draws on a combination of research methodologies (e.g., quantitative analysis of large surveys, structured interviews, machine learning-based models, etc.) and requires active participation from multiple stakeholders (e.g., program planning and administration, data analysts, education researchers, etc.). We believe that future research is necessary to establish performance benchmarks for financial literacy interventions and help guide policy making.

References

Abreu, M., & Mendes, V. (2010). Financial literacy and portfolio diversification. *Quantitative Finance, 10*(5), 515–528. https://doi.org/10.1080/14697680902878105

Akerlof, G. A., & Shiller, R. J. (2010). Animal spirits: How human psychology drives the economy, and why it matters for global capitalism. In *Animal spirits: How human psychology drives the economy, and why it matters for global capitalism*. https://doi.org/10.5455/ey.20020

Almeida, F., Duarte Santos, J., & Augusto Monteiro, J. (2020). The challenges and opportunities in the digitalization of companies in a post-COVID-19 world. *IEEE Engineering Management Review*, *48*(3). https://doi.org/10.1109/EMR.2020.3013206

Aren, S., & Dinç Aydemir, S. (2014). A literature review on financial literacy. *Finansal Araştırmalar ve Çalışmalar Dergisi*, *6*(11). https://doi.org/10.14784/jfrs.2014117326

Asaad, C. T. (2015). Financial literacy and financial behavior: assessing knowledge and confidence. *Financial Services Review*, *24*(2), 101–118.

Atkinson, A., & Messy, F.-A. (2012). *Measuring financial literacy: Results of the OECD infe pilot study* (OECD Working Papers on Finance, Insurance and Private Pensions, 15).

Bajo, E., Barbi, M., & Sandri, S. (2015). Financial literacy, households' investment behavior, and risk propensity. *Journal of Financial Management, Markets and Institutions*, *III*(1), 157–174. https://doi.org/10.12831/80534

Bannier, C. E., & Neubert, M. (2016). Gender differences in financial risk taking: The role of financial literacy and risk tolerance. *Economics Letters*, *145*. https://doi.org/10.1016/j.econlet.2016.05.033

Barnett, M. L., Mehrotra, A., & Landon, B. E. (2020, April 29). Covid-19 and the Upcoming Financial Crisis in Health Care. *NEJM Catalyst*.

Bayer, P. J., Bernheim, B. D., & Scholz, J. K. (2009). The effects of financial education in the workplace: Evidence from a survey of employers. *Economic Inquiry*, *47*(4). https://doi.org/10.1111/j.1465-7295.2008.00156.x

Bernheim, D. (1998). Financial illiteracy, education and retirement saving. In *Living with defined contribution pensions* (Issue 1996).

Blue, L. E., & Grootenboer, P. (2019). A praxis approach to financial literacy education. *Journal of Curriculum Studies*, *51*(5). https://doi.org/10.1080/00220272.2019.1650115

Braunstein, S., & Welch, C. (2002). *Financial literacy: An overview of practice, research, and policy* (Federal Reserve Bulletin, 88). https://heinonline.org/HOL/Page?handle=hein.journals/fedred88&id=1303&div=&collection=

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1). https://doi.org/10.1037/0022-3514.42.1.116

Calvet, L. E., Campbell, J. Y., & Sodini, P. (2009). Measuring the financial sophistication of households. *American Economic Review*, *99*(2). https://doi.org/10.1257/aer.99.2.393

Cavezzali, E., Gardenal, G., & Rigoni, U. (2012, October). *Risk taking, diversification behavior and financial literacy of individual investors* (Working Paper). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2162783

Cole, S., Sampson, T., & Zia, B. (2011). Prices or knowledge? What drives demand for financial services in emerging markets? *Journal of Finance*, *66*(6). https://doi.org/10.1111/j.1540-6261.2011.01696.x

Collingwood, J. (2016). *What's your time perspective?* Healthline Media. https://psychcentral.com/lib/whats-your-time-perspective#1

Diacon, S. (2004). Investment risk perceptions: Do consumers and advisers agree? *International Journal of Bank Marketing*, *22*(3). https://doi.org/10.1108/02652320410530304

Dvorak, T., & Hanley, H. (2010). Financial literacy and the design of retirement plans. *Journal of Socio-Economics*, *39*(6). https://doi.org/10.1016/j.socec.2010.06.013

Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A. Y., Foufou, S., & Bouras, A. (2014). A survey of clustering algorithms for big data: Taxonomy and empirical analysis. *IEEE Transactions on Emerging Topics in Computing, 2*(3), 267–279.

Faraj, S., Renno, W., & Bhardwaj, A. (2021). Unto the breach: What the COVID-19 pandemic exposes about digitalization. *Information and Organization*, *31*(1). https://doi.org/10.1016/j.infoandorg.2021.100337

Fox, J. J., & Bartholomae, S. (2008). Financial education and program evaluation. In *Handbook of consumer finance research*. https://doi.org/10.1007/978-0-387-75734-6_4

Gale, W. G., & Levine, R. (2013). Financial literacy: What works? How could it be more effective? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2316933

Glass, D. C., Kogan, N., & Wallach, M. A. (1965). Risk taking: A study in cognition and personality. *The American Journal of Psychology, 78*(3), 516. https://doi.org/10.2307/1420596

Greaves, D. (2019). Making sense of big data using cluster analysis. *Impact, 2019*(1), 25–29.

Guiso, L., & Jappelli, T. (2008). *Financial literacy and portfolio diversification* (Working Paper). European University Institute. https://cadmus.eui.eu//handle/1814/9811
Hathaway, I., & Khatiwada, S. (2008). *Do financial education programs work?* (Working Paper 08-03). Federal Reserve Bank of Cleveland.
Holzmann, R. (2010). Bringing financial literacy and education to low and middle income countries. In *Bringing financial literacy and education to low and middle income countries*. https://doi.org/10.1596/27828
Hung, A., Parker, A. M., & Yoong, J. (2011). Defining and measuring financial literacy. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1498674
Huston, S. J. (2010). Measuring financial literacy. *Journal of Consumer Affairs, 44*(2), 296–316. https://doi.org/10.1111/J.1745-6606.2010.01170.X
Jacob, K., Hudson, S., & Bush, M. (2000). Tools for survival: An analysis of financial literacy programs for lower-income families. *Woodstock Institute*, *60605*(312).
Kaiser, T., Lusardi, A., Menkhoff, L., & Urban, C. (2022). Financial education affects financial knowledge and downstream behaviors. *Journal of Financial Economics, 145*(2), 255–272.
Klement, J. (2015). Investor risk profiling: An overview. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2597691
Lin, C. L., Lee, S. H., & Horng, D. J. (2011). The effects of online reviews on purchasing intention: The moderating role of need for cognition. *Social Behavior and Personality*, *39*(1). https://doi.org/10.2224/sbp.2011.39.1.71
Lusardi, A., Michaud, P. C., & Mitchell, O. S. (2017). Optimal financial knowledge and wealth inequality. *Journal of Political Economy*, *125*(2). https://doi.org/10.1086/690950
Lusardi, A., & Mitchell, O. S. (2007). Baby Boomer retirement security: The roles of planning, financial literacy, and housing wealth. *Journal of Monetary Economics*, *54*(1). https://doi.org/10.1016/j.jmoneco.2006.12.001
Lusardi, A., & Mitchell, O. S. (2014). The economic importance of financial literacy: Theory and evidence. *Journal of Economic Literature*, *52*(1). https://doi.org/10.1257/jel.52.1.5
Lusardi, A., & Mitchelli, O. (2007). Financial literacy and retirement preparedness: Evidence and implications for financial education. *Business Economics*, *42*(1). https://doi.org/10.2145/20070104
Lusardi, A., Schneider, D., & Tufano, P. (2011). *Financially fragile households: Evidence and implications* (Brookings Papers on Economic Activity, 1). https://doi.org/10.1353/eca.2011.0002
Lyons, A., & Kass-Hanna, J. (2021). A multidimensional approach to defining and measuring financial literacy in the digital age. *SSRN Electronic Journal*. https://doi.org/10.2139/SSRN.3873386

Michie, S., Atkins, L., & Gainforth, H. L. (2016). Changing behaviour to improve clinical practice and policy. In *Novos desafios, novas competências: Contributos Atuais da Psicologia*. https://doi.org/10.17990/axi/2016_9789726972679_041

Moore, D. (2003). *Survey of financial literacy in Washington State: Knowledge, behavior, attitudes, and experiences*. Washington State University (Issue SESRC Technical Report).

Müller, S., & Weber, M. (2010). Financial literacy and mutual fund investments: Who buys actively managed funds? *Schmalenbach Business Review*, *62*(2). https://doi.org/10.1007/bf03396802

OECD International Network on Financial Education. (2011). *Measuring financial literacy: Questionnaire and guidance notes for conducting an internationally comparable survey of financial literacy*. OECD.

Osberg, T. M. (1987). The convergent and discriminant validity of the need for cognition scale. *Journal of Personality Assessment*, *51*(3). https://doi.org/10.1207/s15327752jpa5103_11

Remund, D. L. (2010). Financial literacy explicated: The case for a clearer definition in an increasingly complex economy. *Journal of Consumer Affairs*, *44*(2). https://doi.org/10.1111/j.1745-6606.2010.01169.x

van Rooij, M., Lusardi, A., & Alessie, R. (2011). Financial literacy and stock market participation. *Journal of Financial Economics*, *101*(2). https://doi.org/10.1016/j.jfineco.2011.03.006

Wang, M., Keller, C., & Siegrist, M. (2011). The less you know, the more you are afraid of-A survey on risk perceptions of investment products. *Journal of Behavioral Finance*, *12*(1). https://doi.org/10.1080/15427560.2011.548760

Xu, L., & Zia, B. (2012). *Financial literacy around the world: An overview of the evidence with practical suggestions for the way forward* (Policy Research Working Paper, 6107).

Zhang, J. W., Howell, R. T., & Stolarski, M. (2013). Comparing three methods to measure a balanced time perspective: The relationship between a balanced time perspective and subjective well-being. *Journal of Happiness Studies*, *14*(1). https://doi.org/10.1007/s10902-012-9322-x

Zimbardo, P., & Boyd, J. (2008). The time paradox: The new psychology of time that will change your life. *Speculative Frontiers*, *1*.

Zimbardo, P., Clements, N., & Rego Leite, U. (2017). Time perspective and financial health: To improve financial health, traditional financial literacy skills are not sufficient. Understanding your time perspective is critical. In *Time perspective: Theory and practice*. https://doi.org/10.1057/978-1-137-60191-9_2

Zimbardo, P. G., & Boyd, J. N. (1999). Putting time in perspective: A valid, reliable individual-differences metric. *Journal of Personality and Social Psychology*, *77*(6). https://doi.org/10.1037/0022-3514.77.6.1271

Zsótér, B. (2018). Examining the financial literacy of young adults. The correlations of time perspective, financial well-being and delay of gratification. *Public Finance Quarterly*, *63*(1).

Zuckerman, M. (2000). *Are you a risk taker?* Psychology Today Singapore. https://www.psychologytoday.com/sg/articles/200011/are-you-risk-taker

Index

T. Walker et al. (eds.), *Big Data in Finance*,
https://doi.org/10.1007/978-3-031-12240-8