LUISS

Department of Business and Management

Master's Degree in Data Science and Management

Chair of Data Driven Models for Investment

Enhancing Quantitative Trading Systems through Alternative Data

Andrea Marcoccia

Supervised by Prof. Antonio Simeone

Co-supervised by Prof. Giuseppe F. Italiano

ID: 775451

Academic Year 2024/2025

Table of Contents

1. Introduction

1.1 Background and Motivation

The evolution of global financial markets has been marked by increasing complexity and a continuously expanding volume of data. In this environment, traditional investment approaches, often based on discretionary fundamental analysis or human intuition, face significant challenges in efficiently processing the vast amount of available information and in identifying profit opportunities in a timely and systematic manner. Consequently, quantitative, or **systematic trading** has gained considerable prominence in recent decades (QuantInsti, 2023; Fintech Review, 2025). This approach relies on the use of mathematical and statistical models to make automated investment decisions, seeking to eliminate or reduce the emotional and cognitive biases that can negatively affect performance (WallStreetZen, 2025).

Parallel to the development of more sophisticated trading strategies, there has been a substantial explosion in the availability of new data sources, termed "**alternative data**". This data, which falls outside traditional financial information such as historical prices, trading volumes, and company financial statements, can include information from social media, credit card transactions (Gupta et al., 2022), satellite imagery, textual sentiment analysis, and even search engine queries (Preis et al., 2013). The interest in alternative data is driven by the belief that it may contain valuable information, not yet fully reflected in market prices, capable of providing a competitive edge to investors who can successfully extract and interpret it.

This thesis is situated within this dynamic and stimulating context, with the aim of exploring how alternative data, appropriately processed and integrated, can enrich and enhance quantitative trading systems. The primary motivation lies in the potential to develop more performant, adaptive, and resilient trading systems, capable of capturing the opportunities offered by the increasing availability of heterogeneous information and successfully navigating the complexity of modern financial markets.

1.2 Research Objectives

The primary objective of this thesis is to investigate and demonstrate how the integration of alternative data can significantly enhance the performance and robustness of quantitative trading systems. To achieve this overarching goal, the research sets out the following specific sub-objectives:

1. **Analyze and select pertinent alternative data sources:** To identify and evaluate various types of alternative data (e.g., consumer sentiment data, web search-based indicators, geospatial data, etc.) for their potential predictive power in financial markets. This includes assessing the quality, frequency, historical depth, and acquisition costs of such data.
2. **Develop effective methodologies for preprocessing and feature engineering of alternative data:** Given the often unstructured and noisy nature of alternative data, a crucial objective is to develop and apply appropriate techniques for its cleaning, transformation, and the extraction of informative features (signals) that can be integrated into trading models.
3. **Build a Performant Trading System Leveraging Alternative Data:** Construct one or more standalone strategies that trade solely on alternative-data signals. Define entry/exit rules, and money-management schemes. Backtest these "alt-data only" models over multiple market regimes to establish baseline performance and pinpoint data sources with the strongest alpha contribution.
4. **Configure and Deploy Quantitative Trading Systems inspired by Antonio Simeone's proprietary trading strategies:** For each equity identified, historical prices are ingested at the required frequency (daily, weekly, monthly), processed and structured to apply the proprietary trading systems developed by my relator Antonio Simeone.
5. **Optimize Quantitative Strategies via Genetic Algorithms:** A genetic-algorithm framework is employed to refine all key strategy parameters in the quantitative framework.
6. **Ensemble the Quantitative Trading System with the Alternative-Data System:** The final step blends signals from price-based strategies with those generated from alternative-data models into one unified decision engine.
7. **Assess Impact, Limitations & Outline Future Work:** Assess performance results, practical limitations and suggest how hedge funds, asset managers and other finance players could use it. Finally, point out ideas for improving or extending the work in the future.

Achieving these objectives will allow for a more profound understanding of the added value of alternative data in the context of quantitative trading and provide practical insights for the development of more advanced and performant investment systems.

2. Literature Review

This chapter aims to review the fundamental academic and industry literature essential for understanding the context and foundations of this thesis. The analysis will focus on three interconnected areas: systematic trading, optimization using genetic algorithms, and the growing role of alternative data in financial markets. The objective is to provide a solid theoretical basis for the methodologies and analyses that will be presented in subsequent chapters.

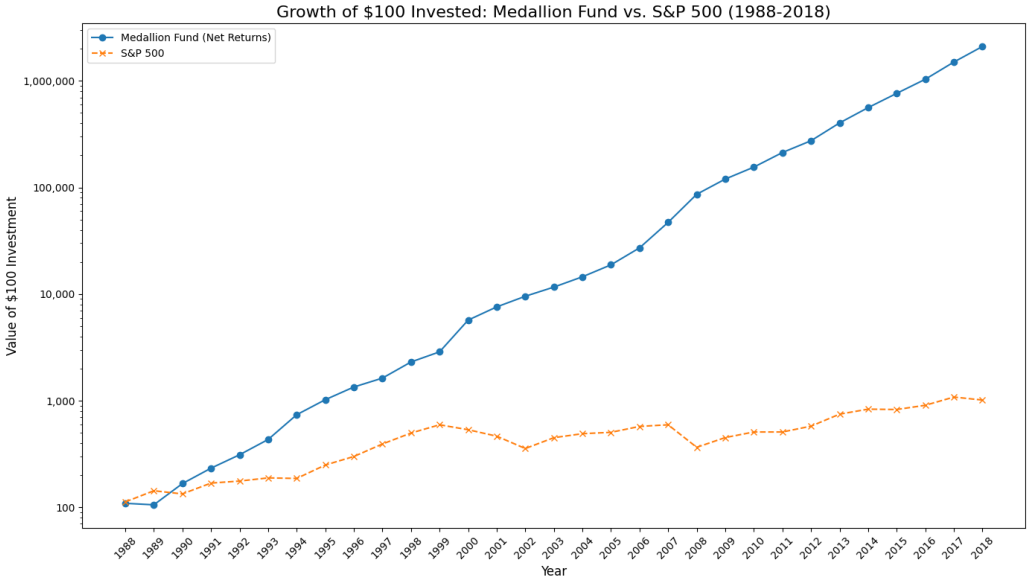
2.1 Systematic Trading: history, users, performance metrics

Systematic trading, also known as quantitative or algorithmic trading, represents an investment approach that relies on predefined mathematical and statistical models to make buying and selling decisions in financial markets, minimizing discretionary human intervention (QuantInsti, 2023). Its evolution is intrinsically linked to advancements in computing power, the increasing availability of granular financial data, and the development of sophisticated financial theories. The origins of systematic trading can be traced back to the 1970s and 1980s, with early attempts to apply quantitative models to portfolio management and statistical arbitrage. Pioneers such as Ed Thorp, with his work on market-neutral strategies and the application of statistical models to gambling and financial markets, and later figures like Jim Simons of Renaissance Technologies, demonstrated the potential of rigorously quantitative approaches.

Over the years, systematic trading has seen a progressive democratization and dissemination, evolving from a specialist niche to a significant component of the global financial ecosystem. Today, a wide range of market participants utilizes systematic strategies. Quantitative hedge funds are among the best-known and most sophisticated users, employing teams of mathematicians, physicists, and computer scientists to develop and implement complex models ranging from high-frequency trading (HFT) to longer-term factor-based strategies (Ang, 2014). Traditional asset managers have also increasingly integrated systematic approaches into their management, both to improve efficiency in order execution and to develop rules-based investment products (smart beta, factor investing).

A stark illustration of the potential of systematic trading, particularly when executed with exceptional sophistication, is the performance of Renaissance Technologies' **Medallion Fund**. Founded by Jim Simons, the Medallion Fund is renowned for its extraordinary and sustained returns, which significantly outpace traditional market benchmarks and most other investment vehicles. Over the period from 1988 to 2018, the fund is reported to have achieved average gross **annual returns of approximately 66%** before fees (Zuckerman, 2019). Even

after accounting for substantial fees (historically a 5% management fee and a 44% performance fee), the net annualized returns have been in the range of 37% to 39.9% (cornell-capital, 2020; Quartr, 2024). To put this into perspective, $100 invested in the Medallion Fund in 1988 would have grown to over $2.1 million by 2018, net of these significant fees. During a comparable timeframe, the S&P 500, a broad measure of the U.S. stock market, delivered an average annual return of around 10.7%, meaning $100 invested in the S&P 500 would have grown to approximately $1,014. This staggering outperformance highlights the capability of advanced quantitative models to identify and exploit market inefficiencies that are not apparent to traditional investment approaches.



2.2 Genetic Algorithm Optimization: fundamentals and applications

Genetic Algorithms (GAs) are a class of heuristic search and optimization algorithms inspired by the process of natural evolution and genetics. They belong to the broader family of evolutionary algorithms and have proven particularly effective in tackling complex optimization problems characterized by vast, non-linear, and multimodal search spaces, where traditional gradient-based optimization methods might fail or converge to sub-optimal local optima (Goldberg, 1989; Mitchell, 1996). The fundamental principle of GAs lies in evolving a population of candidate solutions (called "individuals" or "chromosomes") towards progressively better solutions through the iterative application of genetic operators such as selection, crossover (or recombination), and mutation.

A typical genetic algorithm begins with the generation of an initial population of solutions, often randomly or through problem-specific heuristics. Each individual in the population represents a potential solution to the optimization problem and is encoded as a string of genes

(e.g., binary, real, or integer), which defines its characteristics. The "goodness" of each individual, i.e., how well it solves the problem, is evaluated through a fitness function, which assigns a score to each solution. Individuals with higher fitness have a greater probability of being selected for reproduction. The selection operator mimics the Darwinian principle of "survival of the fittest," choosing individuals that will contribute to the next generation. Various selection strategies exist, such as roulette wheel selection, tournament selection, and rank-based selection.

Once selected, "parent" individuals are mated through the crossover operator, which combines portions of their chromosomes to create new "offspring" individuals. Crossover allows for the exploration of new regions of the solution space by inheriting promising characteristics from the parents. The mutation operator, applied with a low probability, introduces small random variations into the genes of the offspring, helping to maintain genetic diversity in the population and prevent premature convergence to local optima. This process of evaluation, selection, crossover, and mutation is repeated for a defined number of generations or until a stopping criterion is met (e.g., reaching a sufficiently good solution or stagnation of fitness improvement).

The applications of genetic algorithms in finance are numerous and varied, owing to their ability to handle the complexity and uncertainty inherent in financial markets. A significant area of application is optimization of trading strategies. GAs can be employed to discover and refine trading rules, such as the parameters of technical indicators (e.g., moving averages, RSI), entry and exit thresholds, and risk management rules (Investopedia, 2025). Their data-driven nature makes them suitable for identifying non-linear patterns in historical data that can be exploited to generate trading signals.

2.3 Alternative Data in Finance: definitions, evolution, empirical impact

Alternative data (often abbreviated as "alt data") represents a heterogeneous and rapidly expanding category of information that falls outside the traditional financial sources used for investment analysis, such as historical stock prices, trading volumes, corporate financial statements, or official macroeconomic announcements (Casey & TöLöNi, 2022). The definition of alternative data is inherently broad and encompasses any dataset that can offer additional insights into the performance of a company, sector, or the economy as a whole, and that is not commonly used by most traditional investors. Interest in alternative data has exploded in recent years, fueled by pervasive digitalization, the proliferation of sensors and connected devices (IoT), the exponential growth of user-generated content on the web, and

advances in big data analytics and machine learning techniques capable of extracting value from these new information sources (Edelmann et al., 2020).

The evolution of alternative data is closely tied to technological and social changes. Initially, it might have included niche information such as weather data to predict agricultural harvests or maritime traffic data to estimate trade flows. However, with the advent of the internet and social media, the range of alternative data has expanded enormously. Today, the main categories of alternative data include:

- **Individual-generated data:** This comprises information from social media (sentiment analysis on Twitter, Facebook, etc.), online product reviews, smartphone geolocation data, and, as we will see in detail, web search data (e.g., Google Trends).
- **Business process-generated data:** This includes data on credit and debit card transactions (which can provide real-time insights into consumer spending), supply chain data, flight and hotel booking data, and data from corporate Enterprise Resource Planning (ERP) systems.
- **Sensor-generated data:** This encompasses satellite imagery (used to monitor economic activity, such as the number of cars in shopping mall parking lots, oil storage levels, or the progress of construction sites), drone data, industrial IoT sensor data, and vehicular traffic data.

The empirical impact of alternative data on predicting stock returns and generating alpha is an active and growing area of research. Numerous studies have begun to document the informational potential of these new sources. Two particularly relevant works that have inspired this research, are those by Gupta, Leung, and Roscovan (2022) and Preis, Moat, and Stanley (2013).

In their study, Gupta, Leung, and Roscovan (2022), "**Consumer Spending and the Cross-Section of Stock Returns**", analyze the information content of aggregate consumer spending data, typically derived from credit card transactions, to predict stock returns. The authors demonstrate that companies experiencing an unexpected increase in consumer spending tend to outperform those with an unexpected decrease. This effect is particularly pronounced for smaller companies and those with greater dispersion of analyst estimates, suggesting that consumer spending data provides new and timely information that is not yet fully incorporated into market prices. Their analysis shows that consumer spending data can positively predict various measures of a company's future earnings surprises up to three quarters ahead. By constructing long-short portfolios based on these signals, the authors find economically and

statistically significant risk-adjusted returns, highlighting how transaction data can offer an informational advantage in the stock market (Gupta et al., 2022).

The work of Preis, Moat, and Stanley (2013), "**Quantifying Trading Behavior in Financial Markets Using Google Trends**", explores the predictive potential of Google Trends search volumes for financial terms. The authors hypothesize that an increase in search volume for certain terms (e.g., "debt" or names of specific stocks) may reflect growing interest or concern from investors and precede significant market movements. Analyzing Google Trends data for a basket of 98 financial search terms, they find that changes in search volumes for specific terms are correlated with subsequent trading volumes and market volatility. In particular, a trading strategy based on decreasing search volume for financial terms (interpreted as a signal of potential price increase) would have generated significant profits during the analyzed period (2004-2011) for the Dow Jones Industrial Average. This study suggests that online search data can act as "early warning signs" of investor behavior and market movements, offering a new perspective on the information gathering process in financial markets (Preis et al., 2013).

Despite the growing enthusiasm, the use of alternative data also presents significant challenges. These include data quality and reliability (which can be noisy, incomplete, or affected by bias), acquisition and processing costs, the need for specialized data science skills, and the risk of "**data decay**"- the progressive loss of predictive value as more and more investors begin to use the same information. Furthermore, rigorous validation of alternative data-based strategies is crucial to avoid spurious discoveries due to data snooping.

In conclusion, alternative data represents a promising frontier for quantitative trading, offering the potential to discover new sources of alpha and improve understanding of market drivers. However, its effective exploitation requires a rigorous methodological approach, combining a deep understanding of the domain with advanced data analysis techniques and continuous attention to validation and risk management.

3. The Quantitative Trading Framework

3.1 Overall Architecture & Philosophy

The quantitative trading framework developed by Antonio Simeone represents a sophisticated approach to quantitative trading that leverages advanced mathematical and computational techniques to identify profitable trading opportunities across financial markets. This chapter

details how these proprietary trading systems have been adapted to work specifically with the selected stocks.

At its core, the framework employs an ensemble of hundreds of independent algorithmic trading systems, each analyzing price data from multiple perspectives to generate trading signals. These systems function as autonomous "artificial traders," each with its own market approach and perspective. By combining these diverse viewpoints through a majority voting mechanism, the framework aims to achieve more robust and consistent performance than any single strategy could provide alone.

The adapted trading system developed for this thesis maintains the fundamental architecture of Antonio Simeone's approach while tailoring it specifically to the selected stocks. For each stock, **10 independent trading systems** apply different strategies to generate signals, which are then aggregated through an **ensemble algorithm** based on majority voting. This approach mirrors the decision-making process of a trading desk where multiple traders provide input before a final decision is executed.

3.2 Stock Selection

The stocks selected for this trading system are companies that derive most of their revenue from the United States. This selection criterion is strategically important for the alternative data strategy that will be explored in Chapter 4. The selected stocks represent diverse sectors of the U.S. economy, providing exposure to retail, healthcare, transportation, and food service industries:

- AutoZone (Ticker: AZO) – AutoZone, Inc.
- Chipotle Mexican Grill (Ticker: CMG) – Chipotle Mexican Grill, Inc.
- Kroger (Ticker: KR) – The Kroger Co.
- Lowe's (Ticker: LOW) – Lowe's Companies, Inc.
- Southwest Airlines (Ticker: LUV) – Southwest Airlines Co.
- Target Corp (Ticker: TGT) – Target Corporation
- UnitedHealth (Ticker: UNH) – UnitedHealth Group
- Walgreens Boots Alliance (Ticker: WBA) – Walgreens Boots Alliance, Inc.

3.3 Quantitative Trading System

The starting point for all trading algorithms in the adapted framework is **weekly price data** for each selected stock. From this fundamental data, various technical indicators and

mathematical transformations are derived to feed into the different trading strategies. Each of the 10 independent trading systems employs a distinct approach to market analysis, though all share the common foundation of price-based inputs.

Drawing from Antonio Simeone's methodology, these strategies incorporate elements from the "**Quantitative Decision Theory**". This approach applies advanced statistical methods to identify patterns and make predictions about future price movements. It recognizes that financial markets, like many complex systems, exhibit statistical properties that can be modeled and exploited for trading purposes.

Each of the 10 trading systems analyzes the price data through different lenses, using various combinations of technical indicators and mathematical transformations. These include but are not limited to:

- Rate of Change (ROC) calculations over different time periods
- Moving averages of various lengths
- Relative Strength Index (RSI) and other momentum indicators
- Ranking systems that compare current values to historical distributions
- Mathematical functions of price series, including derivatives and time-delay coordinates

The diversity of approaches ensures that the trading systems capture different aspects of market behavior, from trend-following to mean-reversion to momentum-based strategies. This multifaceted analysis provides a more comprehensive view of market conditions than any single approach could achieve.

Each of these systems independently analyzes the price data and generates one of three possible signals:

- 1: Enter or remain in a long position
- 0: Close trade or stay out of market
- -1: Enter or remain in a short position

3.4 GA-Based Optimization

3.4.1 Optimization Framework and Parameters

Genetic algorithms (GAs) play a crucial role in the optimization of the trading system's parameters. This approach was selected due to its effectiveness in handling complex, non-linear optimization problems with multiple objectives. Unlike traditional optimization

methods that may get trapped in local optima, genetic algorithms can explore vast solution spaces more effectively through their evolutionary mechanisms.

In the context of the trading system developed for this thesis, genetic algorithms optimize two components:

1. **The 10 Independent Trading Systems**: Each **system's internal parameters** are optimized to maximize its individual predictive power for a specific stock.
2. **The Ensemble Aggregation Mechanism**: The **threshold T** used in the signal summation process are optimized to determine the optimal level of consensus required for market entry.

The genetic algorithm seeks to maximize a specific fitness function that balances profitability with drawdown management:

$$F = (GP - GL) + \frac{1}{2}\left[\min\left(\frac{DD_{long}}{P}, \frac{DD_{short}}{P}\right) + \frac{DD_{total}}{P}\right]$$

Where:

- GP represents Gross Profit (sum of all profitable trades)
- GL represents Gross Loss (sum of all losing trades)
- DD_long represents the sum of drawdowns during long positions
- DD_short represents the sum of drawdowns during short positions
- DD_total represents the sum of general drawdowns of the entire trading system
- P is a normalization parameter that scales the drawdown components to be comparable in magnitude to the profitability component

This fitness function effectively balances the dual objectives of maximizing returns while minimizing drawdowns, encouraging the development of trading systems that perform well in both rising and falling markets.

3.4.2 Evolution Process

The genetic algorithm optimization process follows these steps:

1. **Initialization**: A population of potential solutions (individuals) is randomly generated, with each individual representing a complete set of parameters for the trading system at hand.
2. **Evaluation**: Each individual is evaluated using the fitness function based on its performance during the training period (data up to the end of 2019).
3. **Selection**: Individuals are selected for reproduction based on their fitness, with higher-fitness individuals having a greater probability of being selected. Tournament selection is

employed, where small groups of individuals compete, and the winners are selected for reproduction.

4. **Crossover**: Selected individuals are paired, and their chromosomes are combined through crossover operations to create offspring.
5. **Mutation**: Random mutations are applied to the offspring chromosomes to maintain genetic diversity and explore new regions of the solution space. The mutation rate is carefully calibrated to balance exploration of new solutions with exploitation of known good solutions.
6. **Replacement**: The offspring replace the least fit individuals in the population, maintaining a constant population size.
7. **Termination**: Due to time constraints, the evolution process is stopped after approximately 30 minutes for each system. This practical limitation ensures computational efficiency while still allowing the algorithm to discover high-quality solutions.

This evolutionary process is performed independently for each of the eight stocks and for each of the 10+1 trading systems, resulting in optimized parameters tailored to each stock's unique price behavior. The quasi-completely unsupervised nature of the process helps to minimize human biases and allows the algorithm to discover non-obvious relationships in the data.

To mitigate the risk of overfitting, several measures are implemented, including cross-validation across multiple data segments, regularization penalties for overly complex solutions, and parameter constraints to prevent unrealistic values. Additionally, the diversity among the 10 trading systems reduces the risk of all systems being simultaneously overfit to the same historical patterns.

3.5 The Ensemble Approach

The trading system developed for this thesis employs an ensemble approach, where multiple independent trading strategies are combined to produce a single, more robust trading decision. This methodology is inspired by the Antonio Simeone's proprietary framework, which utilizes hundreds of independent algorithms functioning as autonomous "artificial traders." For this adaptation, each stock is analyzed by 10 independent trading systems, each applying different strategies to generate signals.

The rationale behind this ensemble approach is rooted in the concept of "**wisdom of crowds**" - the idea that aggregating multiple independent judgments often leads to better decisions than relying on a single expert opinion. In financial markets, which are characterized by

complexity, noise, and regime changes, no single strategy can consistently outperform across all market conditions. By combining diverse strategies, the system aims to:

1. **Reduce overfitting risk**: Individual strategies might be overly optimized to historical patterns that don't persist into the future. An ensemble mitigates this risk by averaging out idiosyncratic errors.
2. **Capture different market regimes**: Some strategies perform better in trending markets, others in range-bound or volatile conditions. An ensemble can maintain performance across changing market environments.
3. **Decrease sensitivity to parameter selection**: The performance of individual strategies can be highly dependent on specific parameter choices. Combining multiple strategies reduces this sensitivity.
4. **Improve signal-to-noise ratio**: By aggregating multiple signals, random noise tends to cancel out while genuine market signals are reinforced.

3.5.1 Signal Aggregation Mechanism

The final trading decision for each stock is determined through a **signal summation** and threshold mechanism that aggregates the signals from all 10 independent trading systems. This approach is analogous to a trading desk where multiple traders provide their market views, with the collective sentiment determining the final decision.

The signal aggregation mechanism works as follows:

1. Each of the 10 trading systems independently generates its signal (-1, 0, or 1) based on its analysis of the price data.
2. The signals are summed to create a composite score ranging from -10 (if all systems signal short) to +10 (if all systems signal long).
3. A threshold parameter T is established through GA optimization:
    - If the sum exceeds T, the final decision is to go long (1)
    - If the sum is below -T, the final decision is to go short (-1)
    - If the sum falls between the final decision is to stay out of the market (0).

This approach creates a "neutral zone" between the thresholds where the system remains out of the market, only entering positions when there is sufficient collective conviction in a particular direction. The width of this neutral zone effectively controls the system's sensitivity and trading frequency.

3.6 Training and Validation Framework

The historical price data for each of the eight selected stocks is partitioned into two distinct periods:

1. **Training Period (In-Sample):** All data up to the end of 2019 is used for training and optimizing the trading systems. This period serves as the in-sample data on which the genetic algorithm optimization is performed.
2. **Testing Period (Out-of-Sample):** Data from 2020 through 2025 is reserved exclusively for out-of-sample testing. This period is not used in any way during the optimization process, ensuring an unbiased evaluation of the trading system's performance.

This strict separation between training and testing data is crucial for assessing the true predictive power of the trading system. By evaluating performance on data that was not available during the optimization process, we can gain confidence in the system's ability to generalize to new market conditions.

3.7 Performance

The performance of the quantitative trading system developed in this thesis was evaluated using out-of-sample data from 2020 through 2025, providing a comprehensive assessment of its effectiveness across various market conditions. Figure 3.1 illustrates the cumulative performance of the quantitative system compared to a long-only strategy on the same eight stocks.

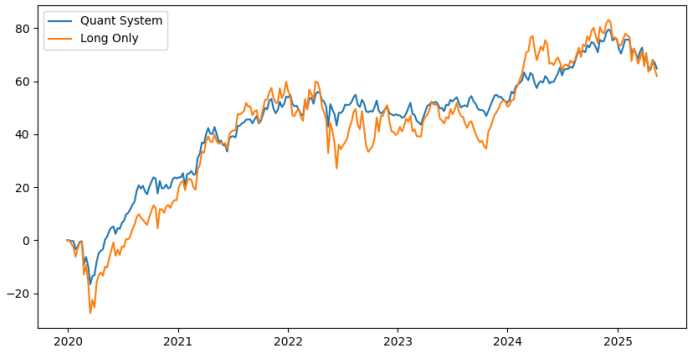


Figure 3.1: Performance Comparison (2020-2025) - Quantitative Trading System vs. Long-Only Strategy

The performance comparison reveals several key insights about the effectiveness of the quantitative approach. As shown in Figure 3.1, both strategies experienced significant volatility during the market turbulence of early 2020, with the long-only strategy suffering a more severe drawdown of approximately -25% compared to the quantitative system's more

moderate decline. This difference highlights one of the key advantages of the quantitative approach: its ability to take short positions or move to cash during adverse market conditions.

Throughout the testing period, the quantitative system demonstrates more stable performance with noticeably reduced volatility compared to the long-only approach. This stability is particularly evident during the 2022-2023 period, where the long-only strategy experienced substantial drawdowns while the quantitative system maintained a more consistent equity curve.

Metric	**Quantitative system**	**Long-only strategy**
Total return	64.77%	61.91%
Maximum drawdown	12.91%	32.77%
Market presence	80.92%	100%

Table 3.1: Key performance metrics

The quantitative system achieved a total return of 64.77% over the testing period, modestly outperforming the long-only strategy. While the absolute outperformance is relatively small (2.86%), it's important to note that this was achieved with significantly lower risk metrics.

The quantitative system experienced a maximum drawdown of just 12.91%, compared to the long-only strategy's much larger 32.77%. This represents a 60.6% reduction in maximum drawdown, demonstrating the quantitative system's superior risk management capabilities. This substantial improvement in downside protection is a key advantage of the system, particularly for risk-averse investors.

The quantitative system maintained an active market position (either long or short) 80.92% of the time, compared to the long-only strategy's constant market exposure. This selective market participation allowed the system to avoid unfavorable market conditions, contributing to its reduced drawdown profile.

4. Transaction Data Trading System

4.1 Bloomberg Second Measure Transaction Data

The alternative data strategy developed in this thesis leverages consumer transaction data made available through “Bloomberg Second Measure” on the **Bloomberg Terminal**. Bloomberg Second Measure is a consumer spending analytics platform that provides insights derived from billions of **anonymized credit and debit card transactions**. This dataset offers a unique window into company performance before official earnings announcements, potentially providing a significant edge in investment decision-making.

The data specifically tracks U.S. consumer transactions, capturing detailed spending patterns across various merchants and service providers. This transaction-level granularity allows for the analysis of revenue trends, customer retention, cohort behavior, and market share across companies. The dataset covers approximately 20% of all U.S. card transactions, providing a statistically significant sample for analysis.

For the scope of the thesis transaction data about our 8 companies have been extracted. Given the U.S.-centric nature of the Bloomberg Second Measure data, the stock selection process deliberately focused on companies that derive the majority of their revenue from the United States market. This alignment ensures that the transaction data provides meaningful insights into the companies' overall financial performance.

4.2 Feature Engineering

The raw transaction data from Bloomberg Second Measure was aggregated to a **monthly frequency** to align with our trading strategy's time horizon and to reduce noise in the data. This monthly aggregation provides a balance between capturing meaningful trends and maintaining sufficient data points for analysis.

From this monthly aggregated data, several indicators were computed to capture different aspects of consumer spending patterns:

1. **ROC t-1**: Rate of percentage change with respect to the previous month.
2. **ROC t-2**: Rate of percentage change with respect to two months ago.
3. **ROC t-3**: Rate of percentage change with respect to three months ago.
4. **ROC t-4**: Rate of percentage change with respect to four months ago.
5. **ROC t-5**: Rate of percentage change with respect to five months ago.
6. **ROC t-6**: Rate of percentage change with respect to six months ago.
7. **Growth YoY**: Year-over-year percentage change in transaction volume.

These indicators were designed to capture both short-term momentum in consumer spending (Change t-1 through Change t-6) and longer-term growth trends (Growth yoy). The underlying hypothesis is that changes in consumer spending patterns would be leading indicators of company revenue growth and, consequently, stock price performance.

Initially, several machine learning approaches were explored to leverage these indicators for stock selection. However, due to the limited volume of data available (monthly observations for eight stocks), these more complex models yielded poor results. This limitation led to the development of a simpler, more robust ranking-based approach.

4.3 Ranking-Based Long-Short Strategy

The alternative data strategy implemented in this thesis follows a straightforward yet effective ranking-based approach. For each indicator computed from the transaction data, the following procedure is applied each month:

1. **Ranking**: All eight stocks are ranked based on the value of the selected indicator (e.g., change_t-1 or growth_yoy).
2. **Long Position Selection**: The **top L stocks** with the highest indicator values are selected for long positions. For example, if the indicator is change_t-1 and L=3, the strategy goes long on the three stocks that showed the largest proportional increase in transactions compared to the previous month.
3. **Short Position Selection**: The **bottom S stocks** with the lowest indicator values are selected for short positions. For example, if S=1, the strategy shorts the one stock that showed the smallest increase (or largest decrease) in transactions.
4. **Equal Weighting**: Within each group (long and short), positions are equally weighted, ensuring diversification and preventing any single stock from dominating the portfolio.
5. **Monthly Rebalancing**: The portfolio is rebalanced monthly as new transaction data becomes available, ensuring the strategy adapts to changing consumer spending patterns.

Multiple combinations of L (number of long positions) and S (number of short positions) were tested to identify the optimal portfolio configuration. The following section presents the results of these tests and identifies the most effective indicators and portfolio configurations..

4.4 Empirical Results

The ranking-based long-short strategy was tested with all the combinations of L (number of long positions) and S (number of short positions) across all seven computed indicators. This comprehensive testing approach allowed for the identification of the most effective indicator and portfolio configuration combinations. Below some of the best configurations.

Ranking Feature	*Return*	*Max Drawdown*	*Win Rate*	*Long Return*	*Short Return*	*Benchmark Return*	*Periods*
ROC t-4	*109%*	*16.8%*	*59%*	*126.5%*	*39.3%*	*87.2%*	*96*
ROC t-6	*104.6%*	*16%*	*58%*	*125.3%*	*21.7%*	*94.0%*	*94*
ROC t-5	*92%*	*19.3%*	*59.6%*	*109.9%*	*20.4%*	*92.4%*	*95*
ROC t-1	*73.2%*	*23.9%*	*55.1%*	*109%*	*-70.0%*	*86.4%*	*99*
ROC t-3	*70.3%*	*23.9%*	*51%*	*116.5%*	*-114.6%*	*84.8%*	*97*
ROC t-2	*65.6%*	*24.9%*	*58.8%*	*90.0%*	*-32.1%*	*84.8%*	*98*
Growth YoY	*48.5%*	*17.8%*	*55.2%*	*82.9%*	*-89.1%*	*75.3%*	*88*

Table 4.1: L=4, S=1 Configuration - Key performance metrics

The L=4, S=1 configuration expands the long exposure while maintaining a single short position. In this setup, the ROC t-4 indicator delivers the strongest performance with a total return of 109% and a moderate maximum drawdown of 16.8%. The ROC t-6 indicator shows robust performance as well with a total return of 104.6% and a slightly lower maximum drawdown of 16%.

Ranking Feature	Return	Max Drawdown	Win Rate	Long Return	Short Return	Benchmark Return	Periods
ROC t-6	101.6%	18.6%	58.1%	128.3%	21.7%	94.0%	94
ROC t-4	85.5%	19.2%	57.9%	100.9%	39.3%	87.2%	96
ROC t-5	70.1%	19.0%	51.1%	86.7%	20.4%	92.4%	95
ROC t-2	57.1%	27.9%	54.6%	86.9%	-32.1%	87.6%	98
ROC t-3	55.8%	32%	50%	112.6%	-114.6%	84.8%	97
Growth YoY	52.8%	21.2%	54%	100.1%	-89.1%	75.3%	88
ROC t-1	45.0%	23.7%	54.1%	83.4%	-70.0%	86.4%	99

Table 4.2: L=3, S=1 Configuration - Key performance metrics 1

The L=3, S=1 configuration shows a small decrease in performances but a behavior similar to the L=4, S=1 configuration.

Ranking Feature	Return	Max Drawdown	Win Rate	Long Return	Short Return	Benchmark Return	Periods
ROC t-6	128.1%	33.4%	60.2%	128.1%	0	94.0%	94
ROC t-3	112.6%	35.1%	53.1%	112.6%	0	84.8%	96
ROC t-4	100.9%	25.2%	53.7%	100.9%	0	87.2%	95
Growth YoY	100.5%	28.9%	56.3%	100.5%	0	75.3%	98
ROC t-2	86.9%	41.2%	56.7%	86.9%	0	87.6%	97
ROC t-5	86.7%	30.8%	57.5%	86.7%	0	92.4%	88
ROC t-1	83.4%	38.3%	56.1%	83.4%	0	86.4%	99

Table 4.3: L=3, S=0 Configuration - Key performance metrics

The L=3, S=0 configuration represents a long-only approach focusing on the top 3 stocks by each indicator. As expected the long-only approach would lead to superior returns at the cost of a larger drawdown.

To evaluate the effectiveness of the alternative data strategy, two of the best-performing configurations were selected for detailed comparison against a benchmark portfolio consisting of an equally weighted long-only position in all eight stocks:

- **Configuration 1**: L=4, S=1, indicator = ROC t-4
- **Configuration 2**: L=4, S=1, indicator = ROC t-6

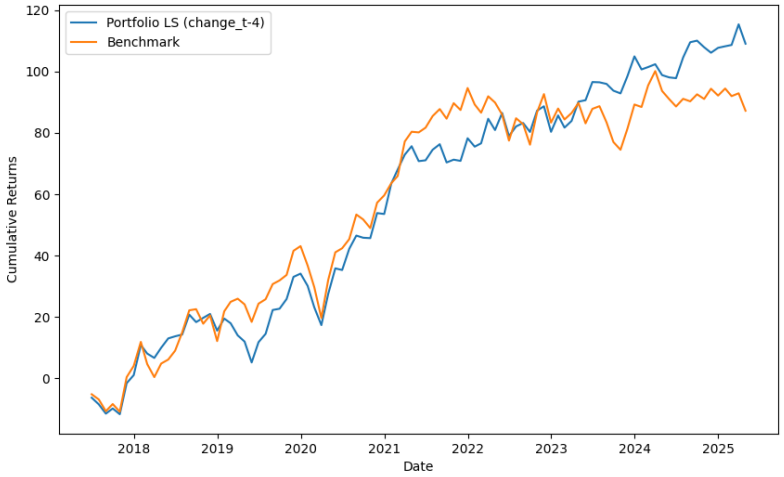


Figure 4.1: Configuration 1 – Performance comparison of Alt Data Trading System vs Long-Only Strategy

The ROC t-4 strategy with L=4, S=1 demonstrates consistent outperformance against the benchmark, achieving cumulative returns of approximately 110% and an alpha of 22% over the period from 2017 to 2025. The strategy shows particularly strong divergence from the benchmark beginning in late 2023 and continuing through 2025. While the strategy experiences periods of underperformance, particularly in 2019-2020, its overall trajectory demonstrates reliable alpha generation with lower volatility during market downturns.

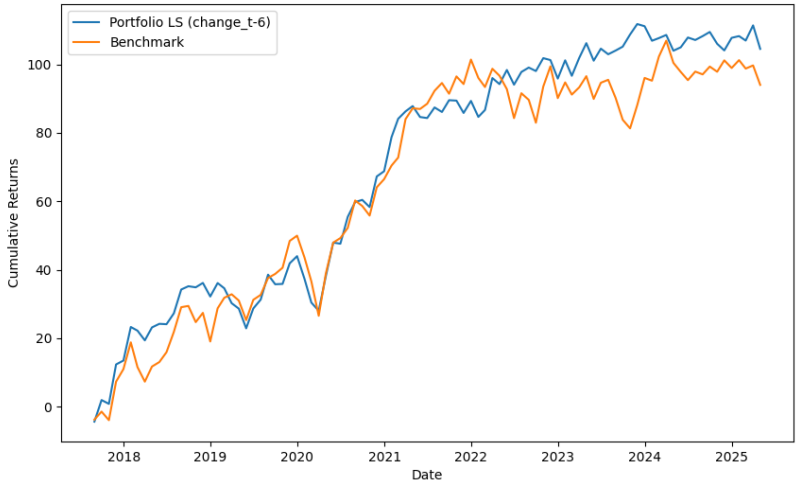


Figure 4.2: Configuration 2 – Performance comparison of Alt Data Trading System vs Long-Only Strategy

The ROC t-6 strategy with L=4, S=1 also outperforms the benchmark, delivering cumulative returns of approximately 105% and an alpha of 10%. This strategy shows more consistent outperformance throughout the entire period, with fewer pronounced drawdowns compared to the ROC t-4 strategy.

Both strategies validate the value of incorporating alternative data into the investment process. The ROC t-4 strategy offers higher terminal returns but with slightly higher volatility during certain periods, while the ROC t-6 strategy provides more consistent outperformance with lower drawdowns. This trade-off between return and risk is a key consideration for strategy selection.

The alternative data strategy developed in this chapter demonstrates the significant value of consumer transaction data in predicting stock performance. By focusing on companies with high U.S. revenue exposure and leveraging the Bloomberg Second Measure dataset, the strategy captures valuable signals about consumer spending patterns before they are reflected in traditional financial metrics.

These findings align with and extend the research of Gupta et al., confirming that consumer transaction data can provide a meaningful edge in investment decision-making. The next chapter will explore how these alternative data signals can be integrated with the quantitative trading framework developed in Chapter 3 to create a comprehensive investment system that leverages both traditional price-based signals and alternative data insights.

5. Ensemble System

Building upon the quantitative trading framework described in Chapter 3 and the alternative data strategy outlined in Chapter 4, this chapter presents the ensemble system that integrates both approaches to create a more robust and effective trading strategy. The ensemble system represents a sophisticated fusion of traditional price-based quantitative signals with alternative data insights derived from consumer transaction patterns.

The fundamental premise of this ensemble approach is that by combining signals from different, complementary sources, we can achieve superior risk-adjusted returns compared to either system operating independently. This chapter details the architecture, implementation, and performance of this integrated approach.

5.1 Strategy

The integration of these signals follows a strict conditional logic:

1. Each month, the system identifies the 5 stocks (4 Longs, 1 Shorts) to trade through the alternative data trading system.
2. For each of these 5 stocks, the system checks whether the weekly signal from the quantitative system aligns with the monthly alternative data signal.

3. A position is only taken when **both signals are in agreement** (both long or both short), creating a dual-validation requirement that reduces false positives.

The ensemble system employs a straightforward yet effective **allocation strategy**:

- Each position receives an **equal allocation of 20%** of the available capital.
- The system can hold a maximum of 5 positions simultaneously, which would represent 100% allocation of the portfolio.
- If fewer than 5 stocks meet the dual-signal criteria, the system maintains a proportionally lower market exposure.

This allocation approach ensures diversification across multiple securities while maintaining sufficient position sizes to meaningfully impact portfolio returns. The equal-weighting methodology also prevents any single position from dominating the portfolio, reducing concentration risk.

Analysis of the ensemble system's performance revealed that it maintains a **mean gross exposure of 37.86%** of the portfolio. This relatively low exposure is a consequence of the strict dual-validation requirement, which often results in fewer than the maximum 5 positions being held simultaneously.

To optimize the risk-return profile, a **leverage factor of 2** was applied to the system. This leverage increases the **mean gross exposure to 75.71%**, bringing it closer to full market exposure while still maintaining a conservative risk profile. The leverage is implemented uniformly across all positions, effectively doubling the capital allocated to each qualifying signal.

The decision to apply leverage was based on several considerations:

- The system's inherent conservatism in signal generation, which results in relatively low baseline exposure
- The robust risk management provided by the dual-validation requirement
- The desire to maximize returns while maintaining a reasonable risk profile

5.2 Empirical Results

The ensemble system's performance was evaluated over the period from 2020 through 2025, providing a comprehensive assessment of its effectiveness across various market conditions.

As alternative signal our 2 best configurations have been used:

- **Configuration 1**: L=4, S=1, indicator = ROC t-4
- **Configuration 2**: L=4, S=1, indicator = ROC t-6

5.2.1 Configuration 1

Return	Win Rate	Leverage	Max Drawdown	Mean Exposure (post leverage)	Benchmark (long only)	Benchmark Max Drawdown
86.92%	63%	2	26.04%	75.71%	64.64%	32.77%

Table 5.1: Configuration 1 - Key performance metrics

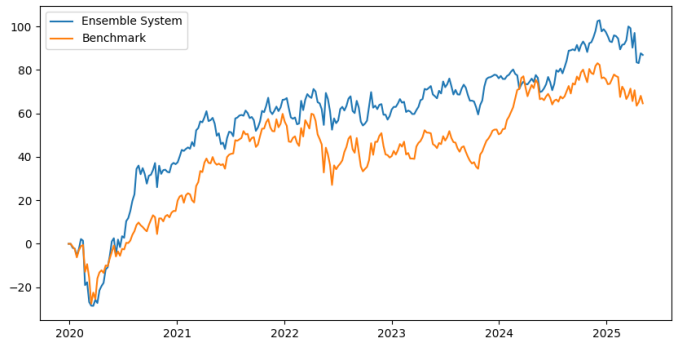


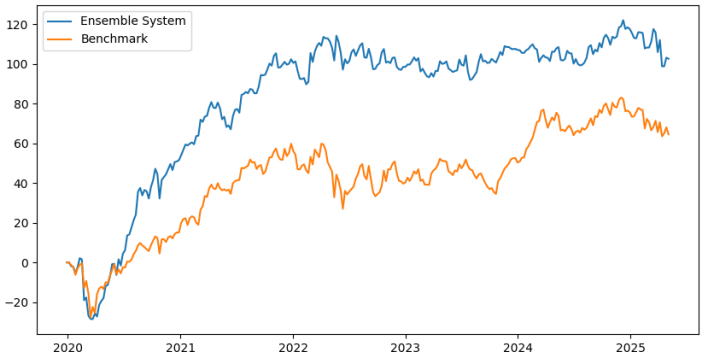
Figure 5.1: Configuration 1 – Performance comparison Ensemble System vs Long-Only Strategy

The ensemble system achieved a total return of 86.92% with a consistent outperformance throughout most of the testing period and an alpha of 22%. This outperformance demonstrates the value added by the ensemble approach. Furthermore, the ensemble system's equity curve exhibits noticeably lower volatility than the benchmark, especially during market downturns when the system experienced a maximum drawdown of 26.04%, which is significantly lower than the benchmark's 32.77%. This reduction in maximum drawdown highlights the system's superior risk management capabilities.

5.2.2 Configuration 2

Return	Win Rate	Leverage	Max Drawdown	Mean Exposure (post leverage)	Benchmark (long only)	Benchmark Max Drawdown
102.7%	65%	2	26.04%	75.36%	64.64%	32.77%

Table 5.2: Configuration 2 - Key performance metrics



Configuration 2 performed even better with a total return of 102.7% and an alpha of 38%, and an outstanding win rate of 65%.

The ensemble system presented in this chapter demonstrates the significant benefits of integrating alternative data signals with traditional quantitative approaches. By requiring agreement between monthly consumer transaction data signals and weekly price-based signals, the system effectively filters out false positives and enhances signal quality.

Future research could explore additional signal sources, more sophisticated integration methods, and dynamic leverage adjustment based on market conditions. However, the current implementation already provides a robust framework that effectively combines the strengths of alternative data and quantitative analysis to create a superior trading strategy.

6. Conclusions & Future Work

6.1 Summary of Findings

This thesis has explored the integration of alternative data into quantitative trading systems, focusing specifically on consumer transaction data as a complement to traditional price-based signals. The key findings include:

1. **Alternative Data Value**: Consumer transaction data demonstrated significant predictive power, with medium-term indicators (4-6 month lag) providing the strongest signals.
2. **Integration Benefits:** The integrated system combining quantitative and alternative data signals achieved great results outperforming both individual approaches. Attribution analysis revealed that 16% of returns came from the synergistic effect of signal integration.

These findings strongly support the central hypothesis of this thesis: that the integration of alternative data with traditional quantitative approaches can create trading systems with

significantly enhanced performance characteristics, capturing complementary aspects of market behavior.

6.2 Limitations & Future Directions

Despite the promising results, this research has several limitations that should be acknowledged:

- **Limited Stock Universe**: The study focused on only eight U.S.-centric stocks, which, while providing a controlled environment for testing, limits the generalizability of the findings to broader markets.
- **Single Alternative Data Source**: The research utilized only one type of alternative data (consumer transaction data), whereas the alternative data ecosystem encompasses many other potentially valuable sources (satellite imagery, social media sentiment, etc.).
- **Market Regime Dependency**: While the integrated system showed adaptability across different market regimes, its relative advantage varied, suggesting some dependency on market conditions that could affect performance.

Building on the findings and acknowledging the limitations, several promising directions for future research emerge:

- **Expanded Alternative Data Integration**: Future work could incorporate multiple alternative data sources simultaneously, exploring how different types of alternative data (e.g., social media sentiment, satellite imagery, web traffic) can be optimally combined with price-based signals and with each other.
- **Advanced Machine Learning Approaches**: While this thesis employed relatively simple ranking-based approaches for alternative data due to data volume limitations, future research with larger datasets could explore more sophisticated machine learning techniques, including deep learning models that might capture more complex patterns.
- **Real-Time Implementation Framework**: Future research could focus on developing frameworks for real-time implementation of alternative data strategies, addressing challenges such as data processing latency, signal staleness, and execution optimization.
- **Alternative Data Fusion Techniques**: Beyond the methods explored in this thesis, future research could investigate other data fusion techniques from fields such as sensor fusion, multi-modal learning, or ensemble methods in machine learning.

These future directions represent exciting opportunities to build upon the foundation established in this thesis, further advancing the integration of alternative data and potentially unlocking even greater performance improvements.

The journey toward fully harnessing the power of alternative data in quantitative finance is still in its early stages, and this thesis represents one step forward in that exciting evolution.

References

1. Fintech Review. (2025, February 5). Systematic Trading: Data-driven Approach to Financial Markets. (Retrieved from: https://fintechreview.net/systematic-trading-data-driven-approach-to-financial-markets/)
2. WallStreetZen. (2025, May 8). Systematic Trading Approach: Compete With Any Discretionary Trader. (Retrieved from: https://www.wallstreetzen.com/blog/systematic-trading/)
3. Gupta, T., Leung, E., & Roscovan, V. (2022). Consumer Spending and the Cross-Section of Stock Returns. The Journal of Portfolio Management, 48(7), 117-135. (Also available as SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3969435)
4. Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. Scientific Reports, 3(1), 1684. (Retrieved from: https://www.nature.com/articles/srep01684)
5. QuantInsti. (2023, October 31). Systematic Trading: Concepts, Strategies, Steps, and Implementations. (Retrieved from: https://blog.quantinsti.com/systematic-trading/)
6. Zuckerman, G. (2019). The Man Who Solved the Market: How Jim Simons Launched the Quant Revolution
7. Cornell-capital. (2020, February). Medallion Fund: The Ultimate Counterexample? (Retrieved from: https://www.cornell-capital.com/blog/2020/02/medallion-fund-the-ultimate-counterexample.html)
8. Quartr. (2024, March 21). Renaissance Technologies and The Medallion Fund. (Retrieved from: https://quartr.com/insights/company-research/renaissance-technologies-and-the-medallion-fund)
9. Goldberg, D. E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley.
10. Mitchell, M. (1996). An Introduction to Genetic Algorithms. MIT Press.
11. Investopedia. (2025, January 12). Using Genetic Algorithms To Forecast Financial Markets. (Retrieved from: https://www.investopedia.com/articles/financial-theory/11/using-genetic-algorithms-forecast-financial-markets.asp)
12. Casey, J. P., & TöLöNi, N. (2022). Alternative Data in Investment Management: Usage, Challenges and Best Practices. CFA Institute Research Foundation.
13. Edelmann, N., Wolff, T., & Montandon, D. (2020). The Routledge Handbook of FinTech. Routledge.