**19 June 2023**

**Group: Big Data Big Dreams**

* Ariq Bintang (22-605-901)
* Luca Gewehr (22-620-967)
* Marco Hafid (22-620-546)

Big Data Analytics – project presentation

**Introduction**

This project aims to determine the optimal quantitative investment strategies considering various investment parameters. We explore a range of sub-questions, from defining the relevant investment parameters to validation of the statistical reliability of the optimal strategy.

Our research uses a collection of data from several sources, including Bloomberg Terminal, World Bank, and Swiss National Bank. The data includes price data of selected indices and currency pairs, Swiss inflation data, CHF money market rates, and spot interest rates on Swiss Confederation bond issues.

We ideated which financial instruments would be most relevant to include in the project and ultimately decided to focus on complementary indices of equities, bonds and commodities, taking into account survivorship bias, hindsight bias and the vast amount of options that would go along with including individual assets.

To find indices with data that is consistent across securities and extends as far back as possible, we investigated data from various sources such as Refinitiv Eikon, Bloomberg Terminal, Wharton Research Data Services and Yahoo Finance.

The methodology involves thorough data cleaning, integration, transformation, and preparation (feature engineering) to ensure the quality of the input for our analysis. We make use of machine learning algorithms to derive optimal investment strategies, with the end goal of this research being not only to uncover these strategies but also to ensure their statistical credibility, making them a reliable tool for decision-making in investment management.

**Research Question**

The main research question is: "What are the optimal, quantitative investment strategies for different possible combinations of investment parameters?" To provide a comprehensive answer, we delve into a set of sub-questions that contribute to the understanding of the factors influencing the choice of optimal investment strategy.

The sub-questions include considerations such as:

1. Identifying the relevant investment parameters that determine the optimal corresponding investment strategy. These parameters could be desired investment objectives, risk constraints, time horizon, future deposits/withdrawals, ESG criteria, asset class restrictions, and geographic restrictions, among others.
2. Choosing the appropriate securities to be considered for the investment universe.
3. Defining the desired criteria for model accuracy and computational efficiency and finding the balance between these two factors.
4. Determining the optimal approach to restricting the possible combinations of securities, ensuring the balance between model accuracy and computational efficiency.
5. Developing a method to determine an optimal investment strategy by comparing equal-length portfolio return series of each candidate investment strategy. This process would incorporate measures such as maximum drawdown, drawdown length, and conditional VaR.
6. Identifying the optimal estimation method and corresponding specification for determining the optimal investment strategy.
7. Validating the robustness and statistical reliability of the optimal investment strategy.
8. Considering the impact of inflation and foreign exchange movements when determining the optimal investment strategy.
9. Evaluating the theoretical underpinnings and assumptions of the optimization model.
10. Identifying potential drawbacks and limitations of the model when applied to real-life investments and finding ways to address or mitigate these issues.

Through this multifaceted approach, we aim to establish a detailed understanding of the optimal quantitative investment strategies based on different possible combinations of investment parameters.

**Data Source(s)**

Our research leverages data from multiple sources, including:

1. "Bloomberg Terminal spreadsheet builder.xlsx" from Bloomberg Terminal, providing price data of selected indices and currency pairs.
2. "API\_FP.CPI.TOTL.ZG\_DS2\_en\_excel\_v2\_5454868.xls" from World Bank Data, providing Swiss inflation data (CPI in %).
3. "snb-chart-data-rendeidglfzch-en-all-20230502\_1430.xlsx" and "snb-chart-data-zimomach-en-all-20230502\_1430.xlsx" from Swiss National Bank Data, providing CHF money market rates and spot interest rates on Swiss Confederation bond issues respectively.

These sources are both reliable and comprehensive, thus well-suited for our research objectives.

For further clarity, the World Bank and Swiss National Bank data files can be found at the following URLs:

* <https://data.worldbank.org/indicator/FP.CPI.TOTL.ZG> (Swiss inflation),
* <https://data.snb.ch/en/topics/ziredev/chart/zimomach> (money market rates),
* <https://data.snb.ch/en/topics/ziredev/chart/rendeidgdtch> (spot interest rates).

**Summary of Methods and Results (review o.b.v. de rest)**

Our methods involve a process of data collection, cleaning, preparation, and analysis, executed in R. We load the data from different documents into R, aligning dates across different datasets, transforming data formats, and performing various operations to derive meaningful metrics and insights. In order to address challenges in data handling and computational efficiency, we've utilized effective data management and efficient R functions and libraries. Also, we've applied machine learning algorithms, specifically xxx and yyy, as part of our analysis to determine the optimal investment strategies. The results derived from this study are then used to identify the optimal investment strategies. The details of the results will be discussed in the 'Results' section. This research is still ongoing, and further specifics will be addressed as we progress.

In addition, details about specific machine learning algorithms should be provided in the "Summary of Methods and Results" section, replacing 'xxx' and 'yyy'.

**Data Collection and Data Storage**

Collecting data was a significant task as it required dealing with a wide variety of sources, each with different data formats. We used specific libraries and functions in R to load data from Excel and CSV files, convert data types, and select the necessary parts of the data. Seeing as the different files include the desired data in different tabs, rows and columns, we had to navigate through this to correctly extract our data.

We kept the raw data in its original formats (Excel and CSV) to ensure the data remained accurate. We used clear and consistent file and folder names to keep track of different versions of the data. For efficiency and easier management, we stored the processed dataframes in a systematic and efficient manner. The raw data was stored in data frames in R due to its compatibility with R's extensive data manipulation and analysis capabilities. These data frames were named descriptively for easy reference and traceability in subsequent stages of the research.

While the raw data is readily accessible for manual review and verification, the manipulated data frames are optimally stored for computational efficiency and the ease of management throughout the subsequent stages of our research.

Data Collection: You've adequately listed the sources and types of data collected. It might be helpful to add the time period for which the data is collected. Is it historical data or updated regularly, for instance, daily, weekly, or monthly? Are there specific dates when data collection started?

**Data Cleaning and Preparation**

Our cleaning and preparation of the data required several key steps. These steps entailed sorting and filtering data, aligning dates across different files to ensure uniformity, and revising column names for better comprehension. We also transformed inflation values into percentages for standardization and computational ease.

Cleaning the data was a pivotal stage in our data management process. We strove for consistency and accuracy, dealing with challenges such as aligning dates from different sources, managing missing values, and standardizing inflation values. To enhance the efficiency of our data cleaning process, we employed the dplyr library's powerful data manipulation functions and used purrr's map functions to implement changes across multiple dataframes.

Cleaning the data was a pivotal stage in our data management process. We strove for consistency and accuracy, dealing with challenges such as aligning dates from different sources, managing missing values, and standardizing inflation values. To enhance the efficiency of our data cleaning process, we employed the dplyr library's powerful data manipulation functions and used purrr's map functions to implement changes across multiple dataframes. Moreover, to improve computational performance, we narrowed down the number of columns, thereby streamlining the process of calculating possible combinations between columns.

In this way, we ensured the data was clean, consistent, and ready for analysis, setting a strong foundation for our research into optimal, quantitative investment strategies.

After cleaning and preparing your data, it's important to validate it before proceeding to the analysis stage. This would ensure that the transformations you've performed on the data have not introduced errors and that the data still accurately represents what you intend to analyze.

Though this may be implied, it may be beneficial to explicitly state the importance of documenting your data cleaning, validation, and analysis processes. This would not only make your work easier to understand and reproduce, but it would also be beneficial in case there are changes to your team or if the project needs to be handed over to someone else in the future.

# Data Analysis and Data Visualization

To conduct our data analysis, we used a variety of statistical and machine learning techniques. We calculated return series in CHF in nominal, real, and excess terms, a task that presented its own challenges due to the sheer volume of data and the need to perform computations over a plethora of securities combinations.

Handling the large volume of data was a considerable challenge, but one that we tackled by using tools specifically designed to deal with such scenarios. The data.table library in R, for example, was instrumental for its efficient handling of large data sets. Parallel processing might have been employed to manage the computational load more effectively.

Our analysis also included finding correlations between daily returns of various indices, implementing quantitative investment strategies, backtesting these strategies, and evaluating their out-of-sample performance.

Data visualization played a vital role in our project, allowing us to present our results graphically. We utilized the ggplot2 library in R for this task due to its robust functionality and effectiveness when dealing with large volumes of data. These visualizations provided us with a clear and concise way to understand and communicate the results of our analysis.

Our data analysis methods were designed to provide clear and concise answers to our research questions, and our visualizations were created to support these findings. As our research progresses, more specifics regarding our methods and their corresponding justifications will be provided, building a comprehensive framework for deriving and evaluating optimal quantitative investment strategies.

The description currently includes a placeholder for the specific machine learning algorithms and statistical methods used. Including details about which specific algorithms and methods were chosen, and the reasoning behind these choices, would provide a clearer picture of your approach and allow others to better understand your analysis. Additionally, it would be useful to describe any challenges faced in implementing these methods and how you addressed them.

**Results**

Our results provide an optimal investment strategy under different combinations of investment parameters. The detailed results will be discussed in a separate document along with tables and figures that illustrate our findings.

r

source("code/result\_generation.R")

The results section is pending completion, given the research is still ongoing. Upon completion, the main findings would be summarized and supported by exhibits (tables or figures), each accompanied by explanatory notes. These exhibits would be generated using R's data visualization libraries.

The results obtained from the analysis, including the optimal asset allocation for different investment parameters and the performance of the proposed investment strategies, will be discussed in the following sections.

Please note that this is a simplified example of the processes involved, and the actual analysis would likely involve more complex steps and techniques.

Our analysis resulted in an optimal investment strategy tailored to different investment parameters. Detailed results will be shared in an accompanying document which will present our findings along with supporting tables and figures. For a glimpse into how we produced these results, refer to our script via source("code/result\_generation.R") in R.

# Scaling and Cloud Deployment

If the project were to be scaled up significantly, we would leverage cloud-based resources for increased computational power and storage capacity. Services such as Amazon S3 could be used for data storage, while Amazon EC2 instances would provide the required computational power. These services offer scalability and are cost-effective. For executing machine learning algorithms on large data sets, we would use Amazon SageMaker, which provides a complete set of tools to build, train, and deploy machine learning models at scale.

Please note that the information provided above and the related code snippets are placeholders and may need modification based on the actual project implementation. Further, the choice of data storage, machine learning models, and cloud solutions would depend on the specific needs and constraints of the project.

To scale up the data pipeline with significantly more data, the use of cloud-based solutions would be considered. Amazon Web Services (AWS), with its wide array of data processing and analysis services, would be a potential choice. The specific AWS services to be used would depend on the exact requirements of the data pipeline, including the need for high-performance computing, machine learning, data storage, and so on. These cloud solutions would offer scalable resources to efficiently handle the increased data volume, ensuring the continuation of the study without compromising performance.

If this project were to expand significantly in scale, we would utilize cloud-based solutions to handle the increased computational demand and storage needs. We might use services like Amazon S3 for data storage and Amazon EC2 for computational power. These services are scalable and cost-effective. If the project requires executing machine learning algorithms on large datasets, Amazon SageMaker would be an ideal choice as it offers a complete set of tools for developing, training, and deploying machine learning models on a large scale.

**Interpretation**

The interpretation section would elaborate on the findings from the data analysis. It would explain the meaning of the results, tying them back to the initial research question: What are the optimal, quantitative investment strategies for different possible combinations of investment parameters?

In this section, the performance of the proposed investment strategies would be compared and contrasted, examining factors such as return, risk, stability over time, and sensitivity to changes in the investment parameters. This analysis would provide insight into the strengths and weaknesses of the different strategies and would indicate which ones might be most suitable for different investment goals and contexts.

To enhance the understanding of the results, this section might also include a discussion of the economic and financial theories or phenomena that underlie the observed patterns in the data. This could include topics such as market efficiency, behavioral finance, and the impact of macroeconomic factors on asset prices.

The interpretation of the results would involve understanding the implications of the identified correlations, the optimal asset allocations, and the performance of the proposed investment strategies. This would also involve considering the limitations of the analysis and the potential areas for further research.

In this section, we would interpret our findings from the data analysis, relating them back to the initial research question: What are the optimal quantitative investment strategies for different possible combinations of investment parameters? Here, we would assess and compare the performance of the proposed investment strategies, considering elements such as returns, risk, stability over time, and how sensitive they are to changes in investment parameters.

**Limitations and Further Research**

Every research study has its limitations and potential areas for further exploration. The following are a few potential limitations and avenues for further research in this study:

**1. Data limitations:** The data used in this study could have limitations such as missing data points, outliers, or inconsistencies. While these issues would be addressed as much as possible during data cleaning, some residual effects might remain. Furthermore, the scope of the data could limit the generalizability of the results. For instance, if the data mainly covers certain regions or periods, the strategies might not perform as well under different circumstances.

**2. Methodological limitations:** The methods used to analyze the data and construct investment strategies could also have certain limitations. For example, they might make assumptions about the distribution of asset returns or the relationships between variables that do not fully hold in reality. Furthermore, the strategies might rely on certain parameters that need to be estimated from the data, introducing the potential for estimation error.

**3. Computational limitations:** The computation required for data analysis and strategy construction could become a bottleneck, especially as the volume of data increases. While parallel processing and cloud computing solutions could be employed to mitigate this issue, they might introduce additional complexities and potential sources of error.

As for further research, this could include extending the scope of the data to cover more regions, periods, or types of assets, exploring alternative methods for strategy construction, or investigating the impact of various other investment parameters. Additionally, more research could be done on the practical aspects of implementing these strategies, such as transaction costs, regulatory considerations, and investor behavior.

One limitation of this study is the assumption of stable correlations over time, which may not hold in reality. Future research could involve exploring dynamic correlations and their impact on investment strategies. Another potential area for further research could be to include more diverse asset classes and to investigate the effect of active versus passive investment strategies.

Every study has its limitations and potential for further research. Here are a few specific to our study:

1. Data limitations: The data used could have missing values, outliers, or inconsistencies that might affect the analysis. Even though we've tried to address these issues during data cleaning, some effects might still be present.
2. Methodological limitations: The methods used to analyze the data and construct investment strategies might have limitations. They could assume certain things about asset returns or relationships between variables that might not be entirely accurate.
3. Computational limitations: The computation required for data analysis and strategy construction might become a bottleneck as the volume of data increases.

Suggestions for further work:

1. Introduction
2. Expand the scope of research to incorporate more data sources for a broader perspective.
3. Introduce additional investment parameters to increase the versatility of the strategies derived.
4. Apply different machine learning models to compare results and enhance the reliability of the optimal strategy.
5. Research question:
6. Explore the influence of macroeconomic factors on the optimal investment strategy.
7. Evaluate the impact of investor behavior and market sentiments on the choice of strategy.
8. Investigate the role of emerging technologies and alternative investments in shaping investment strategies.
9. Data sources:
10. Incorporate data from additional sources to enhance the robustness of analysis.
11. Consider real-time data analysis to account for rapid market changes and shifts.
12. Employ third-party data validation to ensure the accuracy and credibility of the data sources used.
13. Summary of methods and results
14. Use advanced machine learning algorithms to enhance the efficiency and accuracy of the data analysis process.
15. Integrate a feedback mechanism to continuously update the investment strategy based on evolving market conditions.
16. Consider conducting sensitivity analysis to understand the robustness of the derived strategies to changes in various parameters.
17. Data collection, storage, cleaning, preparation, analysis and visualization:
18. Develop a robust data validation mechanism that can verify the integrity and completeness of the collected data from multiple sources.
19. Implement a more sophisticated data cleaning and preparation process using advanced techniques like machine learning-based imputation for missing values.
20. To improve computational efficiency, consider utilizing parallel processing or distributed computing techniques, particularly during data analysis.
21. Expand the data visualization process by incorporating interactive visualization tools like Shiny in R.
22. Consider implementing additional statistical and machine learning algorithms in the analysis process to gain more profound insights and potentially enhance the prediction of optimal investment strategies.
23. Results
24. To validate our results, we could consider a sensitivity analysis, examining how small changes in the input parameters might influence the optimal investment strategy.
25. We should keep an eye on evolving market conditions as these might warrant a modification of the optimal investment strategy.
26. Interpretation
27. Engage financial experts in the interpretation of results to ensure we consider all possible financial phenomena and factors affecting the investment strategies.
28. Contextualize the findings within the broader economic landscape, including current market trends, to ensure the strategies are applicable and valuable.
29. Limitations and further research:
30. Additional research could cover more regions, periods, or types of assets, explore alternative methods for strategy construction, or investigate the impact of other investment parameters.
31. Future research might also consider practical aspects of implementing these strategies, such as transaction costs, regulatory considerations, and investor behavior.