Internet Activity as Alternative Data. The Link Between Web-Search Traffic And Venture Capital Investments.

A Quantitative Analysis

By Erich Ganz

Master Thesis in Business Informatics
Higher School of Economics, Moscow/Russia
Supervisors: PhD. Zeljko Tekic & PhD. Maxim Malyy

Abstract

This master thesis investigates the influence of venture capital on the web search traffic related to the company, subject to investment activity. Time series data of search queries by Google Trends is used to represent web-search traffic. The magnitude of outbreaks in the time series data caused by related investment activity is measured and compared by the types of funding events. Possible structural breaks in the time series data caused by investment activity, are explored. This thesis finds a significant correlation between investment activity and the web-search traffic related to companies. Furthermore, structural breaks are identified, which confirm the theoretical growth dynamics of new ventures. This thesis provides further insights for investors into the dynamics of web-search traffic. It is concluded that Google Trends is valid alternative data, which provides fruitful insights into the investment process.

Contents

1 Introduction

The scientific process is a continuous reshaping of concepts and approaches. Over time these concepts and approaches are reexamined and improved. Improvement might be described as the application of new methods and tools to answer a given question, resulting in a wider understanding of the underlying truth. It is up to researchers explore new tools and investigate how they might be applied for illuminating existing concepts and approaches from a new perspective. In this fashion, this thesis explores the link between web-search traffic and venture capital investment activity. I hope the newly found insights provided by this thesis will help the venture capital and scientific community to improve the understanding of existing concepts and develop new approaches.

The following is an introduction to the field of research, in which this master thesis is embedded. A general framework of knowledge is developed to understand the necessity of research in the given field and the contribution of this thesis to the scientific community.

Companies engage in various types of business activities throughout their lifetime. Activities like marketing and customer service have the goal to improve the public perception and build stable customer relations. On the other hand, activities like accounting and finance yield to control internal processes and efficiently allocate resources. Monitoring those business activities might reveal further information on the current and future condition of a company. This plays a major role in the decision making process for a successful management. In the past, managers relied heavily on experience and various empirical methods in the decision making process. However, time has proven that non-empirical, quantitative methods lead to informed decision making, which deliver significantly better business results.

Those quantitative methods describe the process of evaluating business activity solely based on data, rather than by the application of empirical methods. The quantitative analysis of a company is mainly concerned with modeling and predicting business activity with values. The methods used for such analysis range from traditional econometrics over operations research to modern machine learning algorithms. Companies and their investors nowadays frequently implement quantitative techniques and tools to evaluate the current status of a given company. These quantitative methods require large computational power and data storage. Due to fast development and easy accessibility of computational power and data storage, process and analyzing very large amounts of data has become possible. On the one hand, this trend makes the handling of large data sets reasonably easy. On the other hand, this has created an outstanding demand for more data.

Therefore, the collection of measurable and verifiable data, such as revenues, market share, or investments, is of major concern for any company. The field of business intelligence refers to the procedural and technical infrastructure that aggregates, stores, and analyzes the data produced by the business activities of companies. The goal is to help companies to make informed decisions and to help third parties to gain insights into the performance of the business. Therefore, business intelligence is mainly concerned with delivering valuable and accessible data about a given company.

Business data can essentially be described as the plain facts and statistics collected during the operations of a business. These datasets are likely to be published by the companies themselves, as the information discloses often sensible internal processes or protected interaction with customers. This information generally stays inaccessible to third parties. Therefore, to gain quantitative insights into a given company, external business data analysts heavily rely on the transparency and availability of data. Often this results in a lack of accessibility to necessary data. Privacy concerns, legal issues, or simply the absence of business intelligence insight a company lead to a scarcity of business data. This is a major challenge for business data analysts. Thus, business data analysts seek alternative sources of data, which provide relevant and similar information to a given business and are easy-accessible.

In the area of finance, the predictive capabilities of "alternative data" have been proven. In fields of creditworthiness mobile phone data and social network analysis have significantly added to identify suspicious individuals (see Óskarsdóttir et al., 2019). In stock price prediction alternative data found applications as a temporarily proxy, in periods where no revenue reporting is taking place (see Painter, 2018; Boonpeng and Jeatrakul, 2016). Alternative data has also been successfully implemented in the theoretical formulation of investment strategies (see In et al., 2019). In the investment industry, the venture capital community is known to suffer from data scarcity. This makes it non-trivial to formulate fully non-empirical investment strategies (see Kaplan and Lerner, 2016). Especially young companies, seeking venture capital investments are not able to provide long-term financial records to undermine their investment worthiness. Indeed, investors are known to lack measurable and verifiable data during the investment process and often rely on empirical approaches (see Monika and Sharma, 2015). Therefore, the identification of new, alternative data sources in this area is of high value for the venture capital community. Lately, Malyy et al. (2021) found a significant correlation in the dynamics of web search traffic and external valuation points of technological companies, seeking financing options. Therefore, web search traffic might be considered as an application of alternative data as a proxy for the true value of a young company.

This thesis investigates the influence of investment activity on web-search traffic related to young companies. Firstly, this thesis discusses the investment process in the venture capital industry. It is established why web-search traffic theoretically might serve as a proxy for the true values of a young company. Secondly, web-search traffic time series data is obtained and analyzed. Special attention is devoted to measuring possible outbreaks in web-search traffic caused by investment activity. Possible long-term effects on web-search data caused by investment activity are investigated and discussed. Finally, conclusions are presented on the quality of web-search data as a proxy for young company valuation.

2 The Venture Capital Industry

Venture capital (hereinafter VC) is a form of private equity and a type of financing, which is provided by various types of investors to early-stage companies. When receiving this type of financing, these startups or early-stage companies are referred to as "new ventures". On the one side, there are the new ventures, which are assumed to have the potential to grow fast and reshape the market, in which they operate. On the other side, there are the VC investors, which expect the new ventures to grow fast. The objective of VC investors is to support and accompany new ventures, mostly financially, until the point, where those businesses reach maturity to be no longer be sold. This said, the major difference between VC and common private equity deals is that venture capital tends to focus on young non-public companies, which are expected to grow fast. On the other hand, private equity tends to fund larger and established enterprises, which seek equity or a transfer of ownership stakes. Venture capital fills the gap capital markets and banks leave open due to the high risk associated with a limited operating history and yet develop business models. Even though the gross contribution of VC to the whole economy in the U.S. is assumed to be at 0.8% (see Kjartan Rist, 2020), VC is a major economic driver generating jobs, accelerating innovation, and creating new business models. Vladimirovich et al. (2015) found that VC investments display influence on GDP after four to six years of cross-markets. It is concluded that the effect of VC significantly outperforms traditional investment.

This chapter gives a wide overview of the venture capital industry. Firstly, the different stages of the investment process are described. Companies at the given stages are characterized. Secondly, the role of investors in the venture capital industry is present. Their investment strategies are described, as well as the shortcomings of these strategies.

2.1 Funding Events — The Lifetime Timeline of New Ventures

Venture capital is raised by new ventures at different point its lifetime. There are widely-used, established terms in the VC industry, by which all different funding stages of a new venture are described. This stages play a major role in the theoretical framework of this thesis. Therefore, in the following, these different funding stages are carefully introduced. Firstly, funding events are introduced, which follow time-based order in the lifetime of a new venture. Secondly, other funding events are introduced, which might take place at any point in the lifetime of a new venture. Finally, the finishing event is introduced, where a new venture stops to be privately owned.

Pre-Seed funding is the earliest funding event in the lifetime of a start-up. This stage typically refers to the period, in which the company's founders are just starting to operate the business. Often only a prototype of the core product is on hand. It is yet to get a product on the market. At this stage only few startups manage to raise funds, yet funds provided at this stage are crucial for the initial steps of product development. For a pre-seed startup the most important tasks are market studies, customer profiling and product development — fundamental prerequisites for startup success (see Matt Monday, 2018). Providing capital at the pre-seed stage bears a significant risk as that product may never even make it to market. The well established term in the VC community "FFF" describes the usual landscape of investors at the given stage — Friends, Family and Fools. Therefore, pre-seed investment come with little attention by media and other market participants.

Seed funding is the first official equity funding event. It typically represents the first official money that a startup raises. It might be given to the startup by a legal contract and juridically organization. Usually, this information is made publicly available to the market in the hope to gain the attention of the community. Investment at this stage intends to support the startup to pay for preliminary operations such as market research and further product development. It supports the startup financially until it is ready for further investment. At the given stage startups might have already gained popularity for their ideas and concepts. A substantial amount of enthusiastic customers might be already present. However, the startup lacks the expertise to monetize the business. The fundamental prerequisites (market study, customer profiling, and product development) should be solved already, however, with room for further change. The investor landscape at this point does not significantly differ. It is still mainly formed by FFF. However, investors can also be revenue-based financing lenders, like the local development bank, rich individuals, crowdfunding, early VC investors, or government programs.

The following funding events are referred to as venture rounds. The amounts of capital needed by the startup from now on surpass the capabilities of the former investor landscape. Funding rounds might be performed in different sub-rounds, adding - II, - III, ... to their description. Moreover, the size and structure of the startup might have already reached a stage, where it is more adequate to refer to it as a young company or new venture.

Series A Round funding takes place when all fundamental prerequisites (market study, customer profiling, and product development) have been finally archived. Based on that, the ongoing business operations of the new venture display potential for fast growth. The startup has become a young company and developed a track record, established a customer base, and generates revenue some figures or some other performance indicator. Funds provided at this stage are used to prove the scalability of the core product in the home market. A business plan is to be developed, which will be able to generate long-term profits and incorporate a strategy for expansion to different markets. Moreover, other early-stage business operations are to be performed, like marketing and branding or hiring of high-class specialized professionals. At this stage, the investor landscape changes drastically. A Series A investor often comes in form of a well-established VC firm that manages portfolios of multiple investments in new ventures. Investors at this point often accompany the new venture until it stops being a private enterprise. In the following funding rounds, the investor landscape does not display significant changes.

Series B Round funding is the second funding round for a new venture that has accomplished the fundamental prerequisites. By this time, the new venture has achieved some stability, internal processes are working smoothly, the core product delivers value to customers, and the customer base is growing. Revenues are starting to build up, but this may still not be enough to conquer the home market or start an offensive in a foreign market. Earlier investors have the chance to see how the management team is performing and whether the investment was worth it. Indeed, at this stage, the expectations toward the founders drastically change. The founders stop to just being creative minds and become managers of a multi-million dollar enterprise. That puts extraordinary performance pressure on the founders, as not only the realization of their initial startup idea is dependent on their decisions, but the future of their employees. No later than now the startup might be called a company or new venture. Series B funding rounds happen when a new venture is ready to expand to new markets. Series B funding is there to help to increase its market share in the home market and make the core product scaleable to new markets. Series B investors usually pay a higher share price for their investments in the new venture than Series A investors. This is due to a significant decrease in the risk of such an investment (see Cochrane, 2005).

Series C Round funding takes place when the new venture has a successful working business model, which has been proven to be saleable across markets. The core product or service generates strong demand in their markets and has a substantial and stable customer base. These new ventures look for additional funding to help them develop new products, expand further into new markets, or even to acquire other companies. Series C funding is focused on scaling the new venture, growing as quickly and as successfully as possible. The founders have proven to be successful managers, which are not only able to lead a growing company, but also to profile and sell themselves. This is an important skill as the series C funding round is supposed to be the final boost, and, therefore, attracting investment is crucial. Ideally, the series C funding is the final funding round before the new venture stops being privately owned. Investors from previous funding rounds tend to participate in the Series C funding round as well. This round of funding often attracts new investors as well. Unlike the previous stages of financing, in which most investors are venture capitalists and or rich individuals, large financial institutions such as investment banks and hedge funds seek to participate in the Series C round. This can be explained by the lower risk associated with the investment, since the new venture is already established and successful and because some form of liquidation of the new venture is excepted soon.

Series D, E, F, G, ... Round. The following funding rounds follow a different logic than the previous funding rounds. As mentioned, many new ventures finish raising capital with their Series C funding round. However, there are several reasons why a new venture may choose to go into further funding rounds. Some of which might be positive or negative. Positive reasons might be the following. A new opportunity for growth was identified before stopping being privately owned. The make use of the opportunity more capital is needed. Indeed, new ventures might operate successfully and raise Series D, E, ... funding to even further increase their value before going public. Such a step, however, is taken in accordance with earlier investors, as they tend to hold already large shares in the new venture. Negative reasons for further funding rounds might be the following. The new venture encounters complications in the business plan, which might lead to the necessity for further capital to keep up. For example, the company has not fulfilled the expectations laid out after raising its Series C round. In this case, further funding rounds might be called “down rounds”. Down rounds are perceived negatively by the market and often lead to a lower valuation of the new venture, compared to the valuation, which they reached in their previous round. A down round may help a new venture to make it through complications, however, at the same time, it devalues the stock. After raising a down round, new ventures often find it difficult to raise further capital. This is due to a loss of trust in their ability to deliver on their expectations. However, it is also possible that management and the investors decide to stay private for a long time due to satisfying revenue.

Exit Event. However, if the new venture recovers from the previous down rounds due to complications, it is likely to perform a so-called exit event. An exit event refers to the sale or change in control of a new venture, signaling the "exiting" of private ownership. This is the successful termination of being a private company and the goal of almost all investors and founders. Venture capital investors liquidize their investment in a new venture. The liquidity event can take different forms, an IPO and acquisition being the most common. An initial public offering (IPO) is a public offering, in which shares of the new venture are sold to institutional investors and retail investors. An IPO is typically underwritten by one or more investment banks, who also arrange for the shares to be listed on one or more stock exchanges. Through this process, the privately-held new venture is transformed into a public company. Initial public offerings can be used to raise new equity capital for companies and to monetize the investments of the earlier VC investors or founders. Due to the nature of an IPO, it attracts the attention of the market. However, new ventures reach out to all kinds of companies and advisors to generate more attention, as this results in more interest in the public sold stock (see Michael Peregrine, 2021). An acquisition takes place when the new venture is purchased by another company. Usually, a large number of shares are transferred to gain control of the new venture. The purchased stocks allow the acquirer to make decisions about the new venture without the approval of the founders and early-days investors. Also, this type of exit event causes large attention.

This completes the listing of funding events, which follow time-based order in the lifetime of a new venture. However, a new venture might raise capital by other means at any point in its lifetime. The following describes funding events, which are not linked to a certain order or time.

Loans & Grants. A loan is a financing option for startups and young companies, which are looking to either get started or improve their young companies by some financial institution. On the other hand, grants are usually linked with a government agency that has clear requirements for qualification. Usually, grants are about improving local communities, so companies must be focused on bettering persisting public problems.

Incubator & Accelerator. Incubators focus on early-phase startups that are in the product-development phase and do not have a developed business model. Incubators nurture and mentor startups over longer periods. However, financial support is also to be expected from incubators. A startup accelerator is an organization that offers mentorship, capital, and connections to investors and business partners. It's designed for select startups with promising a minimum viable product (MVP) and founders, as a way to rapidly scale growth.

Angel & Unattributed Venture Capital. Angel Investors are typically high-net-worth individuals who invest very early into the formation of a startup. To be an angel investor, a person does not have to be an accredited investor. Angel investors are wealthy individuals (or groups of wealthy individuals) who invest their own money into companies. This type of VC investment has an individual character. Its terms differ and depend on the so-called "angle and the new venture.

Since angel investors are wealthy individuals, they not only provide capital to a startup but also gain attention from the market. All types of funding events are to some extent desired and designed to attract attention (see Entrepreneur, 2016), which is why new ventures themselves report to the market about events. Every investment by outstanding individuals in the new ventures is proof of trust in the business model. There are various good reasons for the desire of attracting attention to new ventures. Some of which might be the following. Brand visibility, user acquisition, talent and recruitment, and more than all interest from investors and partners. Especial for investors, seeking constantly for new investment opportunities, large media coverage works as bait. The analysis of interest caused in markets by media coverage is subject to this thesis. It will be investigated whether the attention gained by funding events can be translated into web-search traffic. It is of interest, how funding events influence in the short and long term the web-search traffic related to a new venture.

2.2 Investors And Their Valuation Methods

It remains open to describe the intentions and objectives of VC investors. It is of special interest what kind of expertise and skills are required for success in such an uncertain and risky industry. like venture capital. Indeed, the VC industry is risky. Cochrane (2005) finds venture capital to be on average riskier than the S&P 500. However, at the same time he was able to measure the returns from new ventures investments, showing an average return from funding rounds at 72% in early stages (Seed, Series A & B), declining to 46% at later stages (Series C, D, ..), for companies, which completed an exit event. This might explain the attraction of capital despite uncertainty and risk in the VC industry. The capital provided by most VC investors is allocated to high-growth expected new ventures. Those new ventures are usually at some early funding stage. In return for their investment, VC investors receive a portion of ownership of the new venture. If correctly identified and promoted, new ventures start to display high growth. The generated returns, resulting from an exit event, by far outperform the market return (see Cochrane, 2005). The goal of a VC investor is to help the startup to reach a certain maturity level until an exit event can take place. The exit rate describes the number of companies in the portfolio of a VC investor, which successfully exit,

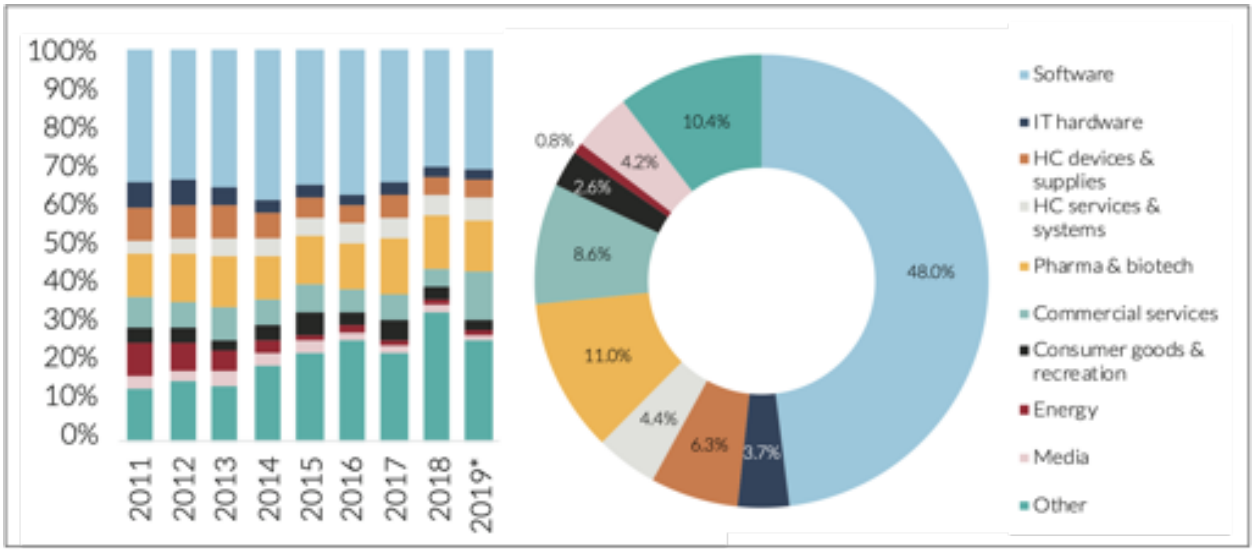

Figure 1: Left — Industry Segmentation of New Ventures From 2011 to 2019 (*until second quarter). Right — Exit Rate of New Ventures Segmented by the Industry in 2019.

compared to the number of companies, which, for some reason, did not reach this point. Therefore, VC investors' major goal is to identify young companies with the potential to grow fast and lead them to an exit event. Therefore, factors must be defined, which serve as indicators for high-growth new ventures.

In their report on VC activity in the US National Venture Capital Association (2019), summarized the industry segmentation over time of new ventures. In figure 1, left graphic, a segmentation is shown, which displays the proportion of different industries, in which new ventures operate. The ten most common markets for new ventures are listed in the legend on the right of figure 1. From this landscape of new ventures, VC investors are to identify potential successful new ventures. The largest group of new ventures operate in "Software". This trend stays consistent over time. Consider the right graphic in figure 3. Shown are exit rates of new ventures grouped by the industry, in which they operate in. It can be observed that the exit rate among "Software" significantly outperforms other industries. Note, that National Venture Capital Association (2019) defines "Software" as a: "Incorporating AI technologies in their core product". This type of new venture forms a large subgroup while displaying simultaneously the largest exit rate. It is beyond the scope of this thesis to provide a durable explanation for this phenomenon. However, an educated guess might take into account the scalability of products in the AI & software industry. Scalability is the property of a company to manage a growing amount of work by adding a relatively small

amount of resources. Scalability might also be described as the ability to quickly access markets at low cost — low entry barriers. In general, this ability is one of the major drivers for economic growth among new ventures (Stampfl et al., 2013, see).

Naturally, VC investors prefer technology-based new ventures, as, by definition, their core product is likely to be scaleable. Technology-based new ventures (hereinafter TBNVs) are new ventures with a core product that is subject to an extensive application of novel technologies. This is done to create and deliver value to customers and stakeholders. Technology-based new ventures create and integrate new technologies into their products and services. For this, new business models based on tech affordance are developed. The major goal is to achieve growth. Technology-based new ventures are not startups. TBNVs started as startups. In their lifecycles, they searched and found the scalable business model, which was further developed to the stage of a new venture.

From the point of view of a VC investor, a portfolio consisting only of TBNVs is therefore likely to display a high exit rate. The total amount of companies in a portfolio of a VC investor depends on whether the investor is an institutional or a wealthy individual (angel investor). For example, according to Sergio Paluch (2021) the most successful angel investor in 2021 was Marc Andreessen, holding investments in 37 companies with 27 successful exits — an exit rate of 73%. On the other hand, the largest institutional investor is Tiger Global (see Vishal Persaud , 2021), holding currently a portfolio of 335 investments. However, the outstanding success of the leading VC investors remains not trivial to explain.

Factors like market size, market competitiveness, the core product itself, and small marketing or revenues play a minor role in the investment process. Venture capital investors are particularly interested in investing in the founders, not focusing exclusively on profit or sales, which are often non-existent for early-stage companies. This makes the valuation of new ventures more art than science. However, research is conducted in this field, though, most of the findings are based on empirical observations. For example, Miloud et al. (2012) identifies three factors, which significantly positively affect the valuation of a new venture by VC investors. The first is the attractiveness of the industry. This factor has been discussed by the example of the tech & software industry in the previous sub-section. However, the quality of the founder and top management team, as well as external relationships of a new venture are considered to be driving factors for evaluating a new venture. Another crucial factor was identified by (Lechler, 2001; Sandberg and Hofer, 1987) as the social interaction inside an early-stage company. The Researchers claim that social interaction can be seen as a source of measurable innovation. Other studies point out the excess of empirical research in this field due to circumstances like the lack of financial history. This creates the necessity for more quantitative approaches, however, adjusted to the given challenges. So proposes Vara

(2013) a systematic risk-based new venture valuation technique, which is an adjustment of "Discounted Cash Flow". This approach is based on estimating the point in time when the risk of offering credit to the new venture with no collateral is minimized. At this point, a DCF is performed and the result is adjusted with comparable new ventures. However, even though most of these studies make use of quantitative methods, the data, which is subject to the research tend likewise to be empirical. A large majority of research in entrepreneurship considers exclusively empirical factors.

The major challenge for the valuation of new ventures is the lack of data. Companies at this stage lack a long-term record of performance, which is why it is not trivial to find trustworthy sources of information/data for evaluation. As new ventures are unable to provide such data, it would need to be provided by another actor. Moreover, as new ventures are expected to grow fast, the data needs to be quickly accessible to the market, for making in-time investment decisions. Finally, this data would need to contain predictive capabilities of the performance of the given new venture. These circumstances display all necessities for alternative data in the valuation process of new ventures. Indeed, lately Malyy et al. (2021) was able to show that web-search traffic might serve as data for growth trajectories of new ventures. The paper finds that the growth dynamics of new ventures are positively correlated with their web-search traffic. However, the question remains open of how this web-search traffic is driven by the actual investment events, which form the fundamental value of the new venture.

3 Web-Search Traffic — A Source of Alternative Data

Due to several developments over the last decade like improvements telecommunication technologies, digitization in our daily life and the worldwide pandemic, a large part of our day to day activities have been moved online (see Johnson, 2021). From online education, stay-at-home health care services or grocery shopping, markets have adjusted their business models to match the new customer needs. One implication of this trend is the generation of enormous amounts of online-data by individuals. This data mirrors preferences, needs and desires of those individuals (see Fessl et al., 2019; Kraut and Burke, 2015). This served as a kick-starter for research making use of this online-data not only on the individual level, but on the group level. Internet activity serves nowadays for modelling dynamics in markets or other social infrastructures (see Pai et al., 2018; Li et al., 2021). This kind of approach makes use of internet activity as "alternative data".

This chapter describes the characteristics of alternative data, on which this thesis conducts a quantitative analysis. An general overview of in the field of alternative data is provided. Web-search traffic is identified as a valid source of alternative data. Current applications of web-search traffic as alternative data are presented and discusses. Finally, a major provider of web-search traffic data is introduced — Google Trends.

3.1 The Uprising of Alternative Data

The term "alternative data" generally refers to a non-traditional datasets used in finance during the investment process (see Forbes, 2019; Wikimedia Foundation, 2022; Fiancial Times, 2020). As shown by earlier examples, this terms might also be encountered in other industries, however, referring to the same concept. In 2016 "Markets Media" published an article, which is today considered as this first consolidation of articles and expert opinions on the term "alternative data" (see Terry Flanagan, 2016). The name of the article, "Early Days For Alternative Data", implies that the term "alternative data" was yet to establish itself in the world of finance in 2016. Although, the core concept of alternative data has been around ever since, a wide usage of this term is observed only nowadays. However, a clear, widely used definition has not emerged by now.

In finance, the investment process considers a subject, in which to invested. The subject can take various types of appearance. E.g. currency, stocks, bonds or raw material. For most of those subjects of investment there exist long-developed, well-established investment strategies. Investment strategies use data to back-up empirical observation and expectations. E.g. the value of a company depends to some degree on the revenue of the company. Therefore, in investment processes predefined sets of variables/data are used.

Three criteria are frequently mentioned, when referring to the term "alternative data" in the investment process (see Raven Pack, 2016; J.P. Morgan Global Quantitative & Derivatives Strategy Team, 2017; Oracle, 2019). The first criterion requires the data to be published not by the subject of investment. If A is the subject of investment then A cannot be a source of alternative data. The second criterion requires the data to provide unique and timely insights into the subject of investment. Unique means that the alternative data, can not be replicated by another data generating process. Timely meaning that the alternative data, is trackable and available over time. The third criterion requires the data to be comparable with data that is traditionally used in the investment process. Therefore, alternative datasets often take the position of proxies for other variables in the investment process, which, on the other hand, are usually published by the subject of investment.

Assume the subject of investment is a company. Then typically alternative data takes the form of a unique and timely dataset, which is not published by the company and mimics some data generating process inside the company. Often alternative data is thought of as a by-product of the operational process of the company, which is not observed by the company itself. Therefore, an alternative description of alternative data is "exhaust data" (see Katherine Noyes, 2016; Techopedia, 2019). The company triggers a data-generating process, which is generally overlooked. Often the company is not aware of the existence of this data generating process and its value to third parties. Since alternative data often originates as a by-product of the company's operations, it is not trivial how to locate and observe it. This leads to complications in the accessibility and/or structure of alternative data (see DUN & BRADSTREET, 2017).

During the last decade, alternative data gained large attention from the financial industry (see Maggio et al., 2022; Jagtiani and Lemieux, 2019; Djeundje et al., 2021). Well-established financial players like J.P. Morgan have included alternative data as a key factor in their overhaul strategy (see J.P. Morgan Global Quantitative & Derivatives Strategy Team, 2017). Consequently, data brokers have emerged, exclusively focusing on providing alternative data as a business model. E.g. in Russia, the largest alternative data broker, Data Fork, provides a wide variety of products like social media sentiment, web-search traffic, or satellite imagery (see Data Fork). The connection to the investment process of such data is not immediately understandable. However, later in this chapter examples will clarify their contribution. The Business Research Company (2021) estimated the size of the alternative data market in 2021 to be \$2.41 billion. This market is expected to grow by 2025 to \$8.98 billion. The report mentions factors like the worldwide pandemic, government-driven digitization, and the implementation of 5G networks as driving factors for the high and promising expectations. Therefore, it is expected that alternative data will play a bigger role in finance and other industries in the coming years. Based on the responses of 28 large investment companies, Greenwich Associates (2018) estimated the current usage and the growth of different kinds of data and information resources used in the financial industry. In the upper graphic of figure 2, the segmentation of the different sources of data used by those investment companies is depicted. Alternative data makes it into the ranking. Therefore, it is one of the top sources. However, it is mostly used as a secondary or tertiary source in the investment process. More than 50% of the asked investment companies do not use alternative data at all. The lower graphic displays the assumed change in sources used in the investment process. 50% expect alternative data to play a larger role in the future. This is more than all other current sources of information display. This outstanding result proves that professionals anticipate a large potential in such data sources.

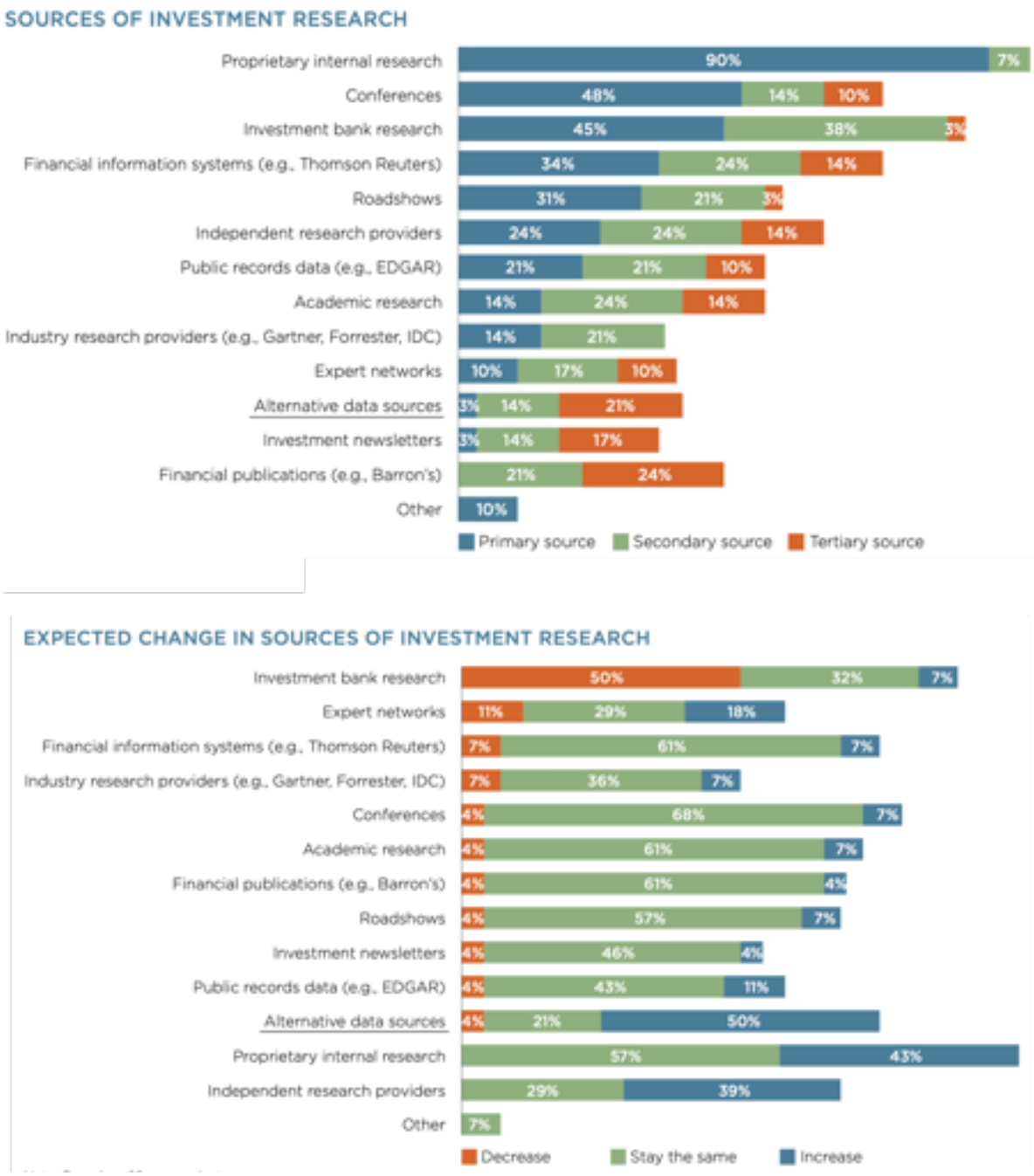

Figure 2: Usage and Expectations Alternative Data (2018)

The question remains of how those investment companies make use of like e.g. satellite imagery in the investment process. Painter (2018) uses satellite imagery as alternative data for stock price prediction. Satellite imagery of parking lots is obtained, which are located in front of stores of a certain retail corporation. With the help of machine learning algorithms, the imagery is converted into quantitative data. The variable of interest is in this case the average number of cars at a firm's retail stores on a given day. The hypothesis claims a measurable influence of the average number of parked cars on the stock's monthly return. A long-short trading strategy based on alternative data is formulated, which earns monthly alphas of 1.6%. It is emphasized that the alpha value persists even after the data is made available to market participants. This leads to the assumption that market participants might not be able to make use of this kind of information due to the non-trivial incorporation of such data into a traditional investment strategy. In another example, Pai et al. (2018) were able to produce satisfying results in stock price forecasting using linear and non-linear combinations of web-search traffic and historical trading data. This hybrid data set of traditional and alternative data provides significantly better predictions.

These examples illustrate that alternative data has become an essential resource for investors in the race for an informational advantage. This type of data continues to drive new research in financial markets as more sources of alternative data are discovered. The earlier cited paper by Malyy et al. (2021) is to be considered as a contribution to the alternative data community. The web-search traffic data used in this paper fulfills the three criteria of alternative data, which were discussed earlier in this chapter. However, unlike satellite imagery, web-search traffic is not only a valuation method for companies. Due to the extremely scarce data circumstance in the VC industry, web-search traffic rather forms the foundation for new research in this field.

3.2 The Predictive Potential of Web-Search Traffic

Web-search traffic might be classified as any action an internet user undertakes for obtaining information. Google takes here a special position. Google is the most popular search engine worldwide and holds in almost all areas, in which it operates, the monopoly for search engine supply. Consumer behavior is being analyzed through online behavior. Take for example the research field "Internet of Behaviors". Internet of Behaviors is an area of research that seeks to understand how, when, and why humans use technology to make purchasing decisions. It combines behavioral science, edge analytics, and the Internet of Things. The Internet of Behaviour refers to the gathering of data that offers important information on client behaviors, interests, and preferences.

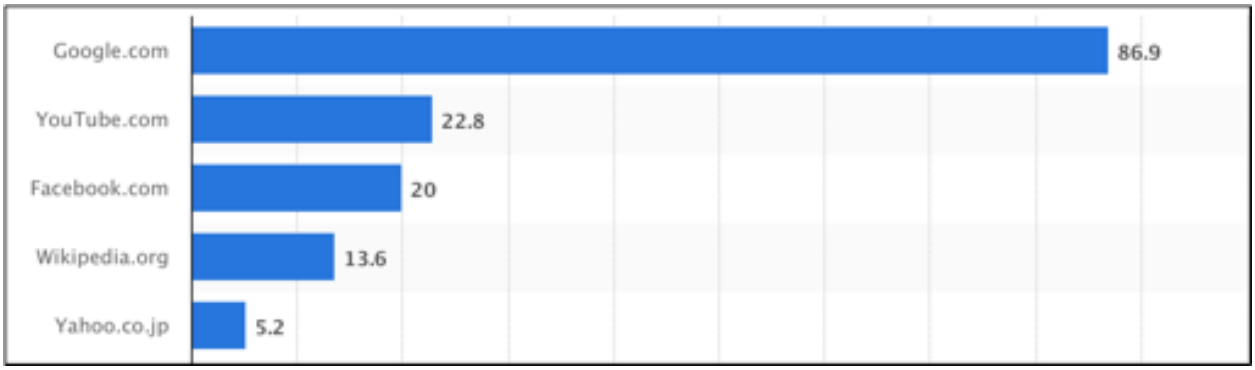

Figure 3: Worldwide Internet Activity in 2021 (First Half)

In this context, one can think of Google as the bottleneck for all kinds of internet activity. To reach out to the world, the first step is the Google search engine. People use the internet to find information, which they don't have. In figure 3 it can be observed that more than 85% of the internet traffic is accounted to google (see DataReportal, 2021). Therefore, taking google's share of web-search traffic might be considered a sufficient analysis of all web-search traffic. Considering figure 3, it is worth recapitulating that the website "Google.com" exclusively provides a search engine. This leads to the assumption that 85% of all internet activity is related to information searches of individuals. These requests for information by individuals can be aggregated into search queries. These aggregates collect information requests into trackable queries. There are many reasons to assume why the aggregation of information requests into search queries might provide certain insights into ongoing processes in all parts of societies. Based on the assumption that the Google search engine serves 85% of all information requests worldwide, this implies a never before seen transparency and accessibility of the dynamics of the societies' interests.

Google provides a tool for the analysis of those search queries — Google Trends. Google Trends (hereinafter GT) search queries can be separated in various regions worldwide and different languages. Google Trends is by now a widely studied source of alternative data. A brief technical introduction to GT is provided in the next chapter. Studies have displayed solid predictable qualities for a wide variety of problems. (CHOI and VARIAN, 2012) give a founded overview of several possible application with the help of GT. However, they mentioned throughout their paper, that the focus of research engaging in GT should rather focus on applications designed for making conclusions on current events, for which current data is not available. They, however, do not exclude that GT might also be a well-suited foundation to build forecasting models. Earlier Varian and Choi (2009) were able to establish a significant relationship between GT data and economic activity. More specifically, it was shown

how queries in the Google search engine, like "Automotive/Vehicle Shopping", significantly correlate with the economic activity of certain areas, where these queries were generated. Last but not least (Malyy et al., 2021) was able to show that GT data significantly correlates with valuation points of a new venture. The afore-mentioned papers have in common, that they relate to GT as a source of data, providing rather information on past activity present, to predict present, unobserved activity, not future activity.

At the end of the previous chapter, the necessity for a new source of alternative data was formulated, with which it will be possible to build valuation models for new ventures, which do not only rely on empirical research. There are several reasons why GT data fits the criteria of such a source of alternative data. Firstly, web-search traffic is not observed by the new venture. It is observed, aggregated, and published by Google. This provides transparency. Secondly, web-search traffic is a unique and timely data-generating process. The period between the generation and the publication of GT data is instant. Other traditional datasets used in the investment process are often made available after a long period. E.g. quarterly financial statements. Mario Gabriele (2021) mentions in his article that: "Venture capital is a game of Speed". Due to the assumed high growth potential, in-time investment decisions are crucial. Thirdly, it has been found that GT data displays similar dynamics as valuation points of new ventures. Therefore GT data mimics to some extent the VC investment process. To research the topic of this thesis, GT data is used as web-search traffic. The short- and long-term effects of investment activity of GT data are explored.

4 Methodology

The previous chapters provided the theoretical foundation for this thesis. Google trends were identified as valid alternative data in the venture capital industry during the investment process. To further understand the dynamics between the investment process and GT data further analysis is needed. This chapter yields to provide the reader with a technical understanding of the instruments used to conduct this analysis Firstly, a technical introduction to GT is provided. Secondly, the business analytics website "CB Insights" is introduced as the source of data on VC investment activity. Finally, details of the data gathering process are presented and conveniences and complications are discussed.

4.1 Google Trends — Technical Details

Google Trends is an interest analytics service for search terms by Google. Google Trends provides time-series data that represent the interest in a given search term over time. Assume there is a finite amount of search terms n. The set of all search terms is Θ. For every search term, τi, there is a subset of search terms Θ_i consisting of all related search terms of τ_j^i.

$$\Theta \;=\; \{\tau_1,\, \tau_2,\, \ldots,\, \tau_n\}$$

$$\forall\, \tau_i \in\; \Theta\; \exists\; \Theta_i \subset\; \Theta,\;\; where\;\; \Theta_i = \{\tau_1^i,\, \tau_2^i,\, \ldots,\, \tau_m^i\}$$

All search terms, τ_i, have an associated time series, ρ_i, which represents how often the given search term has been requested by a user in the Google search engine at any point in time. The interest of a search term is defined by the search query of the topic itself and some proportion, α, of the sum of all related search queries. Note, that the GT documentation does not disclose the exact, α, nor whether α is the same for every related search query.

$$\forall\, \tau_i \in\; \Theta\; \exists\; \rho_i,\;\; where\;\; \rho_i = \begin{bmatrix} \rho_1^i \\ \rho_2^i \\ \vdots \\ \rho_t^i \end{bmatrix}$$

$$f_{interest}(\tau_i) \;=\; \rho_i \;+\; \alpha \sum_{\tau_j^i \in \Theta i} \rho_j$$

A typical search term might be the dairy product "Milk". Image 4 shows how GT provides different types of queries in response to the search term "Milk". When the needed search query was selected, GT provides a list of all related queries to the query for the dairy product "Milk". Every search term gets assigned a code (hereinafter GT code), which relates the search term to its query. Such a GT code usually looks like: "/g/11f017ds55" or "/m/012r5lnd". The exact decision algorithms for determining which queries are related to each other, or how much a related query contributes to the overall interest are not disclosed in the GT documentation. However, further information might be found at https://support.google.com/trends.

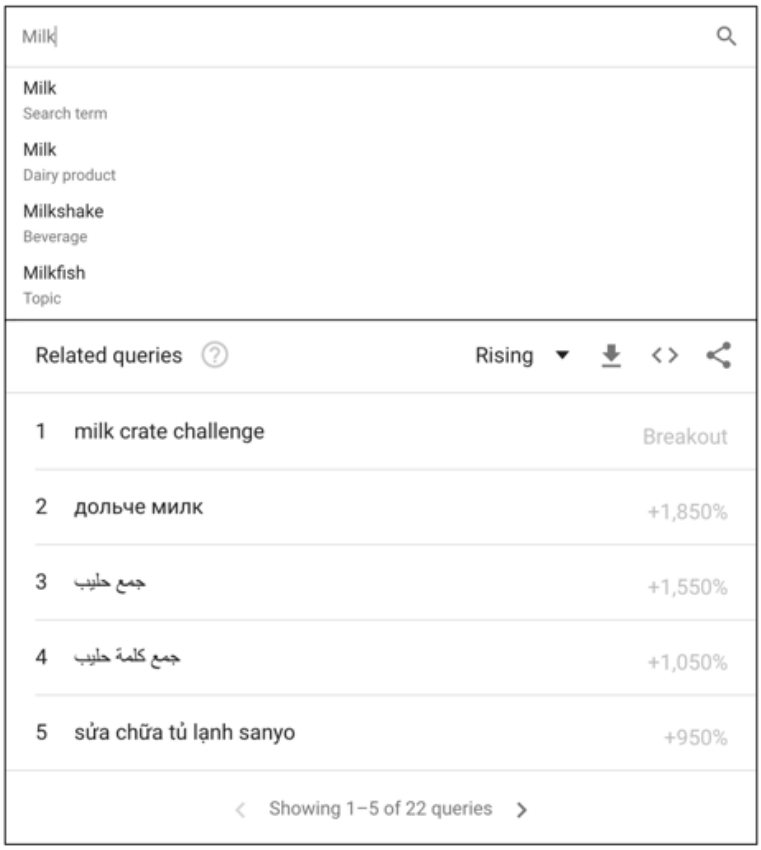

Figure 4: Google Trends - Search Terms

The GT data, representing interest over time, is provided on a daily, weekly, or monthly based. The interest over time is not represented in number of requested. The numerical value of the interest over time ranges from 0 to 100. Therefore, in a given period, the data point with the highest amount of interest is given by 100, while the data point with the fewest interest is given by 0. All other values are proportionally interpolated. Note that in any given period, there is always at least one data point equal to zero, which implies, that this point represents the minimum interest for the period. Assuming that not all data points are equal to 0, this implies that at least one data point must be equal to 100. At this data point, the most interest was registered for the given period. Therefore, only the maximum and minimum interest are observed in GT data, not cardinal amounts of search requests by users. This makes it complicated to compare GT data values from the same or different search terms over different periods and among each other.

Figure 5: The Google Trends Website

An illustration of this problem is to be observed in image 5. Considering the first illustrated graph, representing the interest for the search term "DeepMap" (a new venture) in the period from January 2015 till March 2015, one can observe a maximum value of 100 in the first days of March 2015. This point in time is also illustrated in the second graph. However, the second graph represents the interest for DeepMap in the period from January 2015 till December 2021, which implies that a data point has been added, which now represents the new maximum interest in the given period. Consequentially, this new data point now takes the value of 100 and all other points get assigned a new proportional value for the new maximum point. Considering the GT data point as cardinal values, representing some exact amount of search requests is misleading. The values should rather be understood as nominal values, which are to be interpreted in relation to each other.

Due to the given technical build-up, there are several complications, which arise while identifying and relating the name of a new venture to the correct GT code. The exact method which is used to obtain complete time series data given a name of a new venture is discussed in the following sub-section of this chapter.

4.2 CB Insights — Business Intelligence Platform

CB Insights is a business analytics company, providing platform services and a global database with market intelligence on private companies and investor activity. The platform is focused on private equity, venture capital, investment banking, and angel investing. Moreover, CB Insights provides market analysis in all kinds of business areas. Those analyses might be collections of successful companies operating in a certain industry or reviews and outlooks for certain industries. The services are provided on a paid subscription.

CB Insights publishes a yearly listing of the most promising private companies in the artificial intelligence (hereinafter AI) industry. The yearly listing is published under the name "AI 100". The companies listed in AI 100 were selected by a data-informed process. Some of the parameters/methods used to access a company are patent analytics, company mosaic scores, business relationships, market sizing tools, competitor data, and news trends. To be considered for the AI 100 listing, it is required to be not public (no fulfilled exit event), have raised VC, and have a core product somehow related to tech and AI. Therefore, by the requirements of the selection process, all companies in AI 100 fulfill the necessary criteria to be classified as TBNV. For this thesis, the TBNV in the AI 100 listing from the years 2018, 2019, 2020, and 2021 were analyzed. More listings do not exist. This resulted in a collection of 314 TBNV, due to the fact, that some companies appear in more than one

listing. It was decided to restrict the scope of companies to the four AI 100 listings to have a coherent group of comparable companies, rather than a collection of loosely related or even unrelated companies. This is to reduce possible, nontrivial anomalies, which might arise while comparing results among companies — anomalies, which are due to market-specific circumstances or other circumstances. Moreover, this decision will simplify the data gathering process of the new ventures' GT data. Due to their potential and/or already successful performance, the listed companies in AI 100 can be assumed to have already attracted interest. This interest is likely to have been registered by GT. Indeed, as it will be demonstrated at a later point, there is no guarantee of finding an existing GT query for a corresponding search term, related to some new venture. It might very well be the case that such a query does not exist for various reasons.

4.3 Data Collection

The data subject to this thesis can be divided into two major parts. The first part of the data is obtained from Google Trends. This part contains a collection of time-series data representing the interest over time for the aforementioned TBNVs. The second part of the data is obtained from the CB Insights business analytics platform. This part contains a complete history of all funding events for every TBNV, which was to be found in one of the AI 100 listings. In this subsection, the exact data gathering process of both parts is described. Moreover, complications during this process and their solutions are presented.

4.3.1 The Google Trends Dataset

For the preliminary selection of TBNV, the listings of the business analytics platform were consulted, resulting in a collection of 314 TBNVs. However, obtaining the GT time-series data for 314 TBNV requires further steps. These steps will be described in the following.

Google Trend provides a python library called "pytrends". This library allows to identify the GT code for a certain search term and send an HTTP request to the Google server to obtain the interest over time, related to the identified GT code. There are various other functionalities, like obtaining the number of related queries or focusing on interest in a certain area of the world. However, sending a request and identifying GT codes are the two major steps necessary to obtain the GT time series of interest for a new venture.

Obtaining a GT code works as follows. Take the company name "Affirm". Using the function "suggestions", pytrends returns a collection of possible GT codes in the following manner:

1) mid: /g/11bc5lhx3f, title: 'affirmation', type: Topic
2) mid: /m/04lg06q, title: 'Affirmations', type: New Age
3) mid: /g/11fkl4bl9w, title: 'Affirm', type: Financial company
4) mid: /m/0d4rx, title: 'Affirmative action', type: Topic
5) mid: /g/11h3bgyfwm, title: 'Self-affirmation', type: Topic

In this case, five suggestions are received. Firstly the GT code is given. Secondly, the title is presented. The title represents the search term a user might type into the google search engine. Therefore, the title is the search term, as described in subsection 4.1. Thirdly, the type is given, to which the GT relates. As mentioned in subsection 4.1 there are various types of search terms. Search term types can provide further information on the exact target of the search query and its GT code. In the given case the third option is to be chosen, as, indeed, Affirm is a fintech TBNV, operating as a financial lender of installment loans. Consider, however, the suggestions for the TBNV with the name "The Yes":

1) mid: /m/06299b, title: 'Bryan Danielson', type: American professional wrestler
2) mid: /m/06pxflf, title: 'Yeshiva World News', type: Topic
3) mid: /m/03cwr2b, title: 'YES Network', type: Regional sports network
4) mid: /m/03cwr2b, title: 'Say Yes to the Dress', type: Say Yes to the Dress
5) mid: /g/1q5jb7hnn, title: 'Yesterday Once More', type: Song by Carpenters

There is no suggestion with a title, that matches the name of "The Yes". The types of all suggestions confirm that the suggestions do correspond to the company "The Yes". Another ambiguous case might include several suggestions fitting the search term of interest. The suggestions for the TBNV "insitro" return one query with a fitting title, however, with a type "Topic", which could mean everything. The second suggestion has a matching title and the type is "Company". In such a case the choice of the right GT code is ambiguous. The second suggestion corresponds to the new venture of interest. However, the type "Topic" is the most common type and could denote everything, including the search term of interest. Therefore, the first suggestion might result in a time series, which also contains relevant information related to our new venture. This is due to the fact, that GT is not able to clearly distinguish and unambiguously assign search requests to the correct query without

mistakes. In fact, after analyzing both time series, resulting from both suggestions, it was found that the first GT code delivers better results.

1) mid: /g/11gjx89qyc, title: 'insitro', type: Topic
2) mid: /g/11fhqkn1j6, title: 'Insitro, Inc.', type: Company

The aforementioned complexities were often encountered during the identification process. Moreover, in more than 50% of the times, the name of the new venture led to no suggestions. In such a case there is no possibility to obtain the time series data representing the interest in the given new venture. The GT code was successfully identified for 115 companies from the initial 314 companies. The list of these companies with the corresponding GT code is provided in the appendix.

The corresponding time series were obtained after identifying the correct GT code related of every of the 115 TBNVs. It was decided to consider for every new venture the a time frame from the year of foundation until 31.12.2021. The frequency was decided to be on weekly base. For that a so-called "payload" was build in pytrend. This payload contains information about the region and time frame of the requested interest. Note that different companies have different founding dates, which result in different lengths of time frames. Weekly data is provided by GT only in a time frame of less than five years. Every request with a time frame exceeding five years results in GT automatically returning not weekly data but monthly data. Therefore, companies with an exceeding time frame of five years, require a more involved data gathering approach.

Figure 6 shows the relevant time period of 12 years for a new venture, which was founded in 2010. In order to obtain the weekly GT data related to this new venture, it is necessary to divide the time period into 3 sub-periods. The "payload" in pytrend then separately is constructed for every of the three time periods. Note, the sub-period have intersections of 2 years. This intersections are necessary to calculate a "scaling parameter". Due to the fact, that every request sent to the Google server returns a time series with maximum value 100 and minimum value 0, it is necessary to determine how nominal numbers in different request relate to each other.

Take the last data point in the first request, representing the interest in the time from 26.12.2015 - 31.12.2015. This data point in also presented in the second request. However, their nominal values are likely to no be the same. A simple way to imagine such a case is when the data point for 26.12.2015 - 31.12.2015 represents the point in time with most

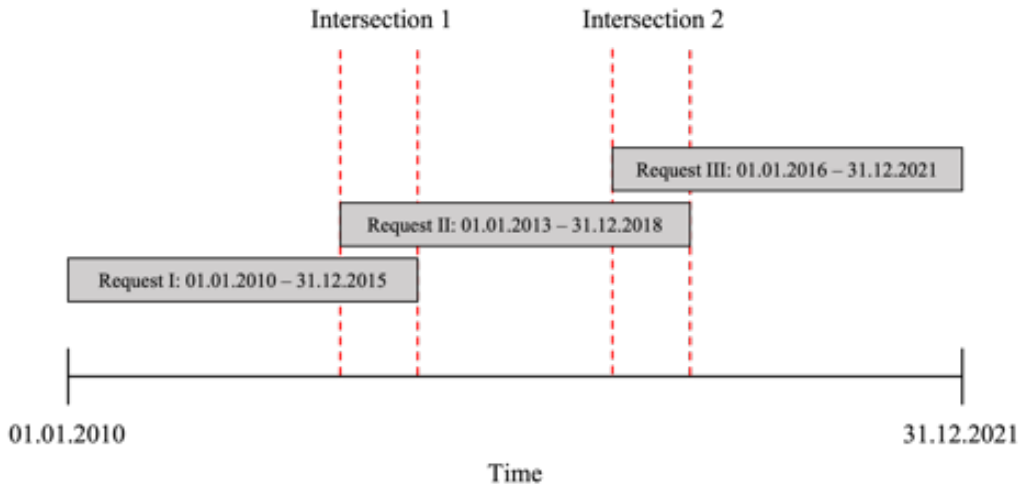

Figure 6: Overlapping Sub-periods For Long Time Periods

recorded interest for the first request. That data point will be assigned a nominal value of 100 in the first request. However, one cannot assume that the same data point will also be the one with the most observed interest in the second request for 26.12.2015 to 31.12.2015. In the second request this data point might will be assigned a nominal value not equal to 100. Therefore, in order to correctly join two requests, it is necessary to calculate the scaling parameter. Afterwards the older sub-period is multiplied by the scaling parameter and both sub-periods are united.

1) $scale = mean(request\,I_{Intersection\,1}) \,/\, mean(request\,II_{Intersection\,1})$

2) $request\,II = request\,II * scale$

3) $request\,I + II = request\,I + request\,II_{afterIntersection1}$

After joining the first and second sub-period, the same approach was applied to join the third sub-period to the ladder two. This approach was applied to all of the 115 companies, which required time periods of more than five years. The result of this approach can be found in figure 7.

Figure 7: Interest Over Time for Company "Algolia"

4.3.2 The Event Dataset from CB Insights

The focus of this thesis is on the affect of a relevant funding event on the we-search traffic, represented by GT data, of a TBNV. CB Insights provides a listings of funding events for companies. One such list is shown in Figure 8. Six variables can be observed. The date of the funding. This will be relevant to assign funding events to their corresponding values in the GT data. The funding round, amount, investors, valuation and number of sources reporting about the given event. Those variables will be used to analyze the different dynamics, resulting from different funding rounds. Note that the funding amounts or valuation are not always disclosed. Therefore, missing data is to be expected. The history of funding has been collected for all 115 companies, for which it was possible to obtain a matching GT code.

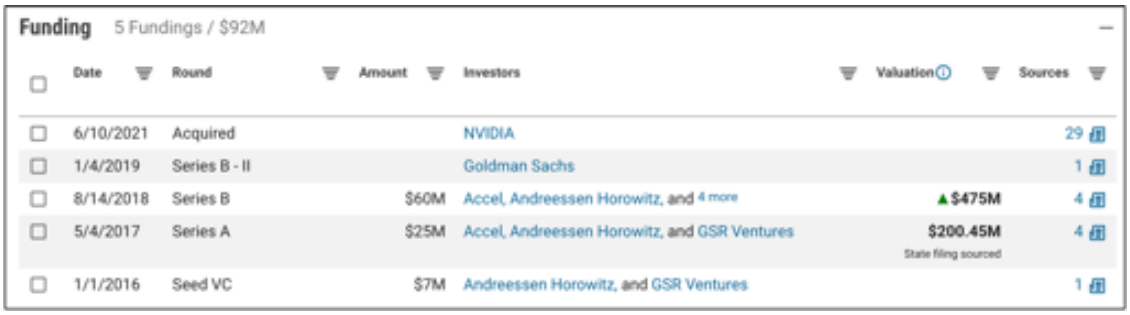

Figure 8: Funding History of Company "DeepMap"

5 Quantitative Analysis

The previous chapter provided an overview of the tools, which were used in this thesis. This chapter presents the quantitative methods and techniques, which were implemented to research the effect of investment activity on web-search traffic. Assumption and their implications are discussed.

5.1 Visual Investigation of the Google Trends Time Series Data

After collecting the GT time series data for the 115 companies, a visual analysis of the resulting graphs was performed. Figure 10 shows four time-series. These four time-series were picked as the display dynamics, which are common in all other 115 companies. They serve as representatives. Five fundamental observations are to be made.

Firstly, the first, third and fourth time series are trended. For the first and third time-series positive trend over time can be observed, which seems to stagnate from a certain point. In terms of interest over time, it is concluded that the interest registered by GT grows over time according to some trends. However, for the last time series, a negative trend is observed, implying a decreasing interest over time for the related search query.

Secondly, comparing the early years of the interest with more recent years, one can observe an increasing fluctuation over. The second and third time-series displays anomaly, while the fourth time series seems to display decreasing variance over time. In terms of interest over time, it is concluded that the variance of the interest is unstable over time and might increase or decrease. Therefore, heteroscedasticity must be assumed among the 115 time-series.

Thirdly, focusing on the data points, which represent interest at the end of every year, one can observe large declines, which seem to deviate from the foregone data generating process. The first and third time-series display this anomaly. The absence of interest, however, cannot be caused by events, which are related to the search term. Every event causing a shock in the data generating process can only cause a positive shock. Even events, propagating negative information related to a new venture, would not display a negative outbreak, but rather a positive outbreak. As already mentioned, these declines are recurring, at a certain point in time. The returning absence of interest at the end of some years might be explained by a seasonal component. Therefore, seasonal components must be assumed among the 115 time series.

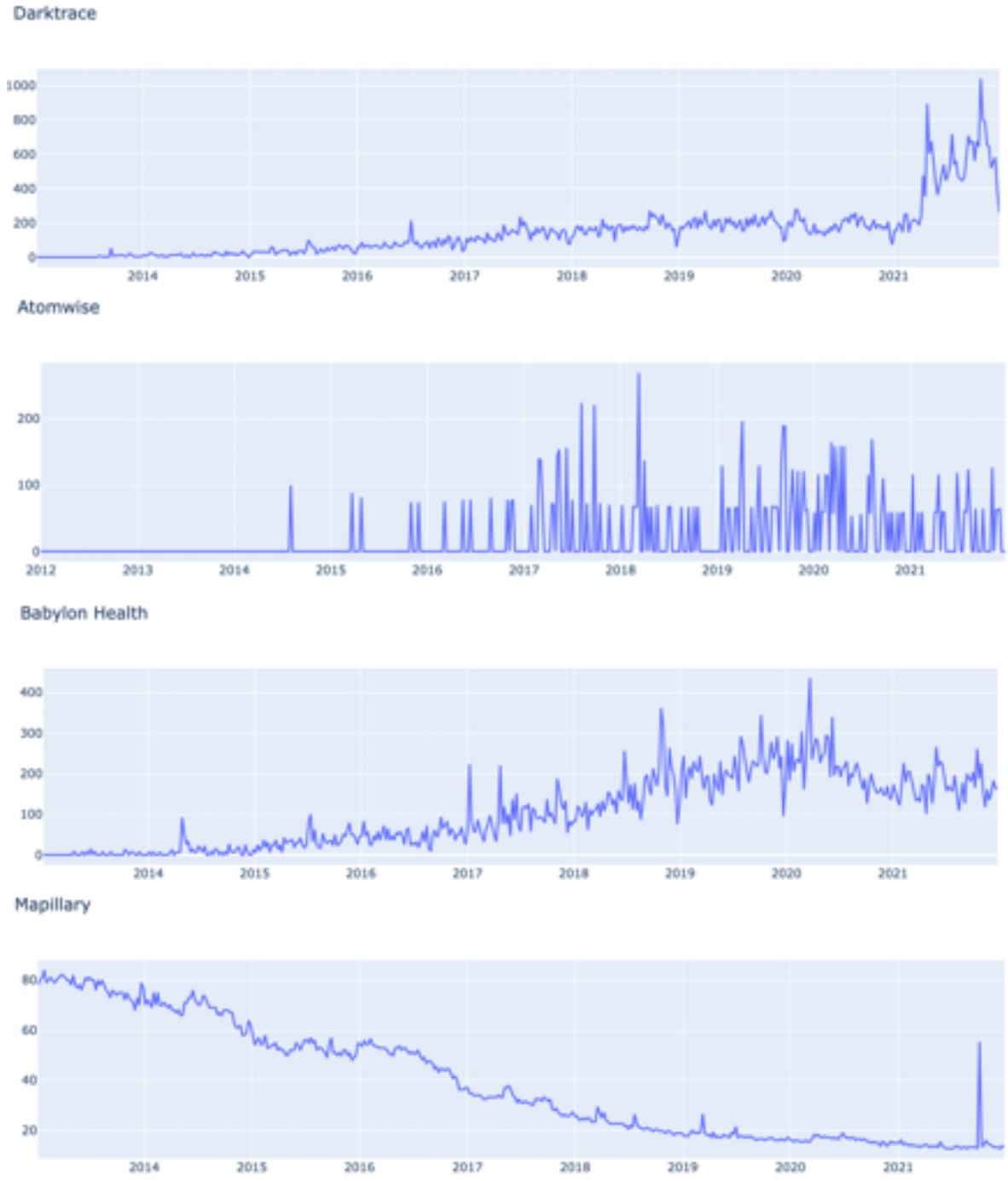

Figure 9: Example Time Series

Fourthly, the last year of the first time-series displays a dynamic, which is not related to the data generating process from the time before 2020. One can observe rapid growth, resulting in a high variance period at a different mean, and then rapidly declining again. In terms of interest over time the presence of events is assumed, which led to a significant temporary long-term increase in mean and variance. Therefore, structural changes must be assumed among the 115 time series.

Finally, all four time series display positive outliers, independent of the point in time. In terms of interest over time, the presence of events is assumed, which causes a temporary shock to the time series. However, it is not clear, whether these shocks cause long-term changes in the data generating process. This implies that outliers must be assumed among the 115 time series.

The visual analysis implies that the resulting time series are not stationary. For later analysis of the dynamics caused by funding events, it will be necessary to identify the underlying data generating process, to measure possible deviations caused by funding events. To isolate this data generating process from the other components, decomposition techniques will be applied.

Before continuing the further analysis, consider in figure 10 the fourth time series. The second and fourth graphs display overall unexpected dynamics. The second graph can hardly be classified as time series. A process is observed, which includes several repeating 0 values with causal, unrelated, and isolated outbreaks. It is not clear, whether this data indeed follows some describable data generating process. The fourth graph displays a declining trend in interest over the whole period. This might very well correspond to the true interest. However, due to the listing of AI 100, one might rather expect increasing interest over time. Therefore, it has to be assumed that some time-series might be subject to mistakes. Reasons for such mistakes in the data can be various. This leads to the question of the quality of the time series data and how this quality might be assessed. This question is not trivial. "What is a good time series?".

Malyy et al. (2021) developed an algorithm for qualification of GT time series data of TBNVs. However, this algorithm has been developed with the priority of determining a meaningful time-series for building regression models and comparing their dynamics to valuation points. The goal of this thesis is not to build regression models for inference, but rather to investigate the dynamics in GT data caused by funding events. For this, it might be sufficient to consider small isolated time frames. Therefore, another method was developed. The details of this method and the exact classification procedure will be described later in this chapter. In the next chapter, the implications of the results of this qualification method will be discussed.

Figure 10: Interest Over Time and Corresponding Events

Now relate the foregone discussion to the given case of this thesis. There are 115 time series, representing interest in new ventures. Additionally, the corresponding funding events for those companies are given. A first step for the analysis of possible effects caused by funding events is to visually inspect the time series at the given points in time. For that consider figure 10. One can observe four time series, where the funding events are highlighted in red. Four conclusions are to be made from this observation.

Firstly, the second time series outliers display a correlation with funding events corresponding to the new venture. Furthermore, one might assume, that there are no further outliers, which are not linked to the funding events. This implies the existence of time series with visible outliers, which might be fully explained by the funding events, corresponding to the new venture. Therefore, it is concluded that the time series data displays some correlation with funding events Secondly, the first and the fourth time series display visible correlation for some funding events. However, for other events, no visual correlation with outliers is observed. Moreover, other data points visually display similar outbreaks, however, without an event having been registered then. Multiple reasons are likely to have caused such dynamics. One could think of significant events, linked to the new venture, however, not connected with investment activity. Thirdly, the third time-series visually does not display any correlation with the corresponding funding events of this new venture. It is of interest why some time-series display high visual responsiveness while others display no visual responsiveness. Referring back to the question of the quality of a time series data, one might assume that the respondence of the GT data to the funding events is independent of the visual shape of the time series. Consider time series 3 and 4. Both display dynamics, unlikely for a common data generating process. However, time-series 4 clearly displays some visual correlation with the foundation events.

5.2 Timeseries Decomposition

In order to further study the underlying data generating process and the influence of funding events, it is necessary to decompose the time series and isolate the data generating process. A first visual inspection has shown a wide range of attributes in the time series data, which need to be accounted for in process of time series decomposition. This sub-section discusses the methods and assumptions, which were applied for the isolation of the underlying data generating process.

5.2.1 Dealing with Outliers

The foregone visual investigation of the time series has shown a visual positive correlation with funding events. Even though such events can cause a visually notable outbreak, it is necessary to formulate a method to quantitatively capture and access those outbreaks.

It is assumed that all given data points might be outliers, caused by funding events. Therefore, it is necessary to correct this outbreaks. It was decided to average all values at the points in time, corresponding with funding events. This averaged values will help to correctly isolate the underlying data generating process. The estimated data generating process will not be influenced by the given outliers. Later in this analysis the predicted value from the estimated data generating process will be compared to the possible outlier at this point. The predicted value at a given funding event serves as an estimator of the interest if the funding event, and therefore the outlier, would have not taken place. For all 115 time series the points in time, corresponding to a funding event, were averaged. Moreover, to ensure that possible dynamics shortly after a shocks are compensated, it was decided to also average the data point related to one week after the event.

The visual analysis has shown more possible outliers, which, however, are not linked to funding events. It was decided to not further identify the exact position or whether they can be indeed be classified as outliers. However, in the following chapter other events are identified, which correlate with this outliers.

5.2.2 Classification of Seasonal Components

The visual analysis has also displayed evidence for a seasonal component in a time series. Seasonal components represent fluctuations over defined time periods that are relatively stable, repetitive and periodical. Factors, by which this recurrent process, is to be determined are timing, direction and magnitude. Identifying whether there is a seasonality component in a time series requires background knowledge about various factor, which potentially influence the time series data. E.g. quarterly performance numbers of a company. However, if such information is not available other methods must be consulted. In the literature the general approach for establishing, whether there is a seasonal component, always includes a graphical approach. The most common graphical approach suggest in this context is a run chart or a run-sequence plot. A run-sequence is a sequence of stack time windows, one for each period in a timeseries. Therefore, the periodically repetitiveness of the time series must be known. The purpose of a run plot is to displays the given data in a time sequence.

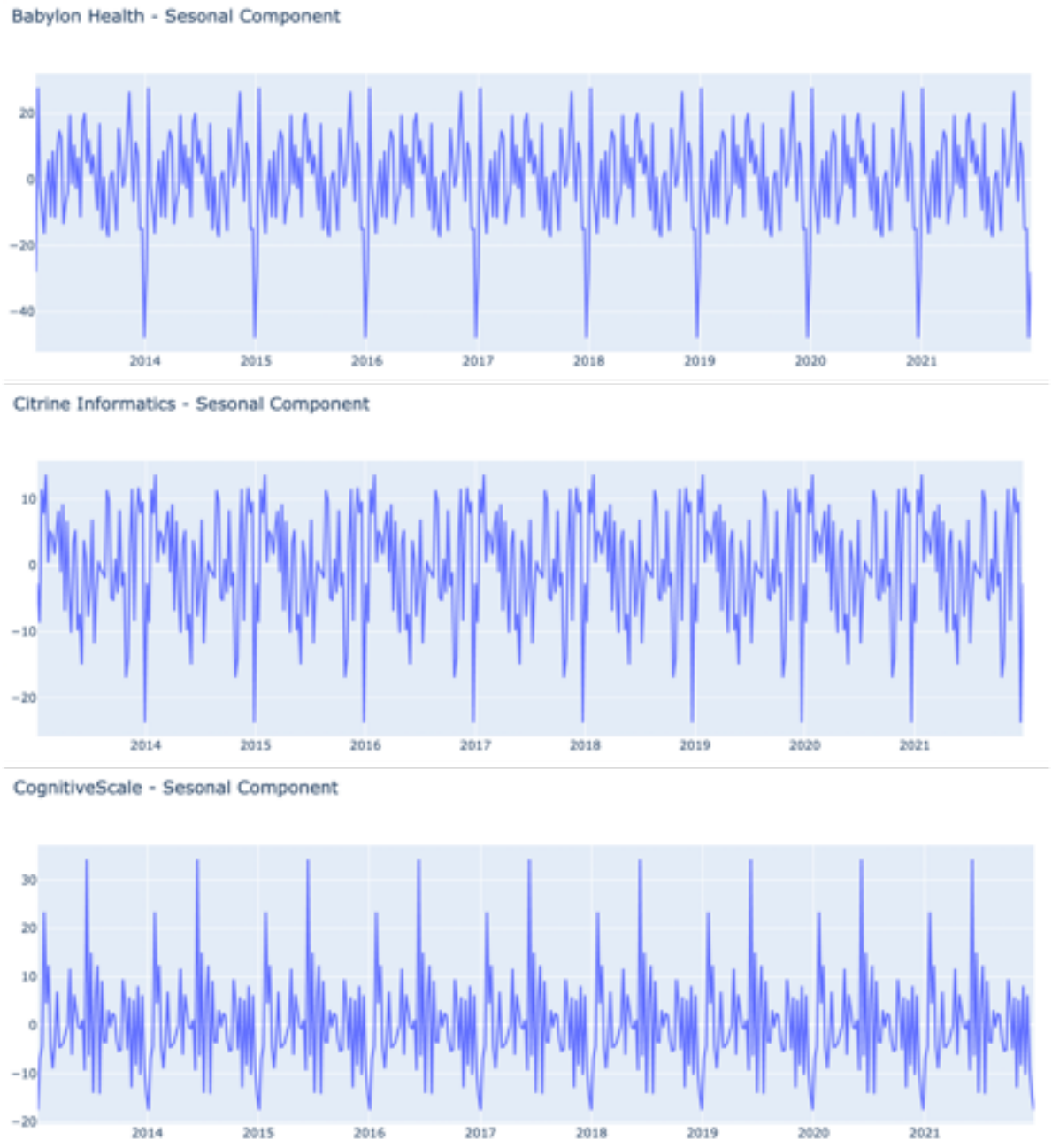

Figure 11: Run-sequence Plots

Figure 11 shows three run-sequence plots, which display representative dynamics for the other time series of all 115 companies. The first run-sequence plot displays a strong negative dynamic at the end of every year. This dynamic has more than twice the magnitude of all other sequential components. The second run-sequence plot displays a clear, however, not as strong seasonal dynamic at the end of every year. The third run-sequence plot displays a rather weak seasonal dynamic at the end of the every year. However the data point at the end of every year does not display the minimum of the whole year. Refer to the description of seasonality as 'relatively stable, repetitive and periodical fluctuations'. The periods in the run-sequence plots, which not correspond to the last week of a year, can no be visually classified as periodical. They more remind of a random walk around some centered value. This is a strong indication for absence of seasonality. Moreover, the given observations at the end of every year not qualify as stable. The observed anomaly is rather an abrupt and short-term dynamic. This raises the question of weather the strong and repeating negative anomaly at the end of every year might be classified as "seasonal".

This type of anomaly is rather common among various search terms and their corresponding GT queries. To illustrate this fact, consider figure 12. Figure 12 shows the interest over time of two common search term, "Screw" and "Microsoft Excel". Putting all other details and dynamics of the time series apart, one can clearly observe an abrupt short-term decline in interest at the end of every year. Below the both time series, the corresponding run-sequence plots are shown. The run-sequence plots clearly confirm the assumption of a negative seasonal component at the end of every year among different search terms and their corresponding queries.

Note in the run-sequence plot of "Microsoft Excel" the short-term appearance of the negative seasonal component. The abrupt negative downfall is recovered within the first two data points after it. There is no further notable repeating dynamics to the process. This is the same anomaly observed among the run-sequence in figure 11. To further establish the existence of the negative anomaly in the seasonal components, consider the run-sequence plot of "Screw". Again a recurring abrupt short-term negative downfall is to be observed at the end of every year. However, run-sequence plot of "Screw" not only displays the negative outbreak. Besides the already described anomaly, clear positive, repeating and stable seasonal component is to be observed. This component appears during the summer months of every year and slowly declines heading towards the end of the year. It is a persistent, slowly increasing and than decreasing dynamic. This can be visually classified as a seasonal component. Note, the given repeating anomaly at the end of every year interrupts the seasonal component. However, it does not influence the underlying dynamic of the seasonal component.

Figure 12: Seasonality in Common Search Terms

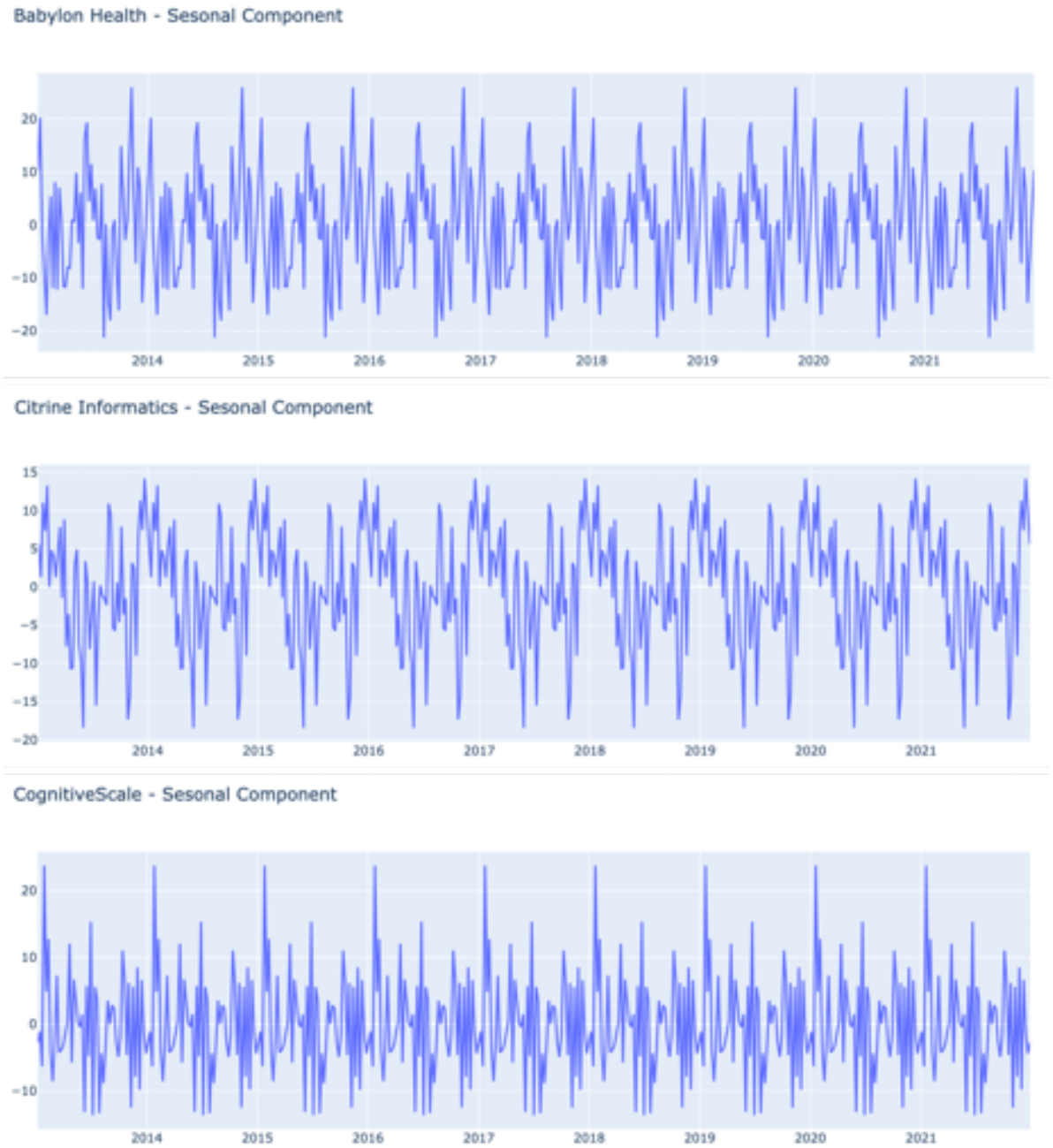

Figure 13: Eliminated Seasonal Component

For dealing with seasonality, common techniques in time series analysis suggest seasonal differentiation. The seasonal difference of a time series is the series of changes from one season to the next. In the given case, for weekly data, the seasonal difference of a time series, y, at period t is $y_t - y_{t-53}$. The value, y_{t-53}, which is subtracted from each value, y_t, of the time-series is simply the value that was observed in the same season one year earlier. Seasonal differencing usually removes the gross features of seasonality from a time series. However, by applying this procedure a large part of the data gets lost. In the given case 53 data points.

Considering the given type of repetitive anomaly in the given time-series data of the 115 companies, another approach is necessary. It was decided to consider the last week of every year as an outlier and average its value by the values of the two data points before and after the last week of the year. For all 115 companies, the last week of every year was averaged by the just mentioned method. Figure 13 shows the resulting run-sequence plots after averaging the last data point of every year. One can observe a dynamic, which visually reminds of a random process. As already mentioned before, this is a reliable indicator of the absence of seasonal components. Moreover, the disappearance of the negative repetitive anomaly can be observed.

5.2.3 Isolation of the Data Generating Process

The dynamics driving the interest of the given companies might have uncountable origins. Considering that the AI 100 listings include companies, active in various industries, the factors are likely to vary among the companies themselves. Identifying exogenous variables, which show significant predictive power for the interest in the given companies is beyond the scope of this thesis. However, for the isolation of the data generating process, it is required to describe the endogenous variable by a set of exogenous variables, which follow the same timeline as the dependent variable. A common approach to overcome the lack of exogenous data is to use past values of the endogenous variable, to predict the present value. Therefore, it was decided to describe all 115 time series by using autoregressive models. Furthermore, the visual inspection of the time series has shown, that the data generating process is subject to a trend and unstable variance over time. To counteract a trend in time series, the general approach is to consider the first difference in the time series. Therefore, it was decided to use an integrated model to describe the different time series. To counteract the unstable variance, the general approach suggested the usage of a process, which takes into account the past variance of the process. Therefore, it was decided to use moving average models to describe the variance in the data generating process.

For the given 115 companies the best fitting autoregressive integrated moving average model (hereinafter ARIMA) model was identified and regressed on the time series data. Classifying the order of the best fitting ARIMA models was done automatically with the auto_arima function in the "pmdarima" library in python. The maximum number of lags and moving average components were chosen to be five and five. The maximum number of integrations was chosen to be one. After having regressed 115 ARIMA models the corresponding residuals of the regressions were analyzed. Note that those residuals are the approximation of the assumed underlying data generating process.

The best-fitting ARIMA model was determined based on the time-series data, which has been cleaned before from its outliers and seasonal anomaly. However, the time series data which is only corrected for the seasonal anomaly is also of interest, as it displays the dynamics caused by funding events. At this point of the analysis, there are two types of residuals, which may be obtained for every time series. The first type of residuals, type-I, results from subtracting the prediction, given by the ARIMA process from the cleaned time series data. The second type of residuals, type-II, results from subtracting the prediction, given by the ARIMA process from the time series data, where only the seasonal anomaly was corrected. Therefore, type-II displays the shocks caused to the underlying data generating process by the funding events. Only deviations from the underlying data generating process are considered. Thus, type-II provides the possibility of comparing outbreaks among different new ventures. Figure 14 shows the time-series graph of the company "Coveo" and the two above-mentioned types of residuals. The red dotted lines and points indicate a funding event, which took place at the corresponding point in time. A strong positive correlation is visually observed between the date of the funding event and the corresponding value in the time series. However, this relation is not displayed in type-I. As mentioned above, this type was obtained with the time-series data, which does not include outbreaks caused by funding events. Therefore, for further analysis type-II is of interest.

5.3 Measuring Outbreaks Related to Funding Events

The fundamental intention of this thesis is to explore the influence of funding events on the interest in a new venture. In the previous chapters, it was mentioned, that any event, related to a new venture must be assumed to cause a positive shock to the data generating process or to not influence the data generating process. Even negative events would theoretically cause an interest in the web-search traffic. To understand this fundamental assumption, it is helpful to consider the meaning of the variable in the time series — namely, interest.

Figure 14: type-I & type-II Residuals

The existence of interest itself assumes the existence of more than one subject of interest and generators of interest, which are endowed with a certain amount of interest. Assume the generators of interest are aware of the existence of some of the subjects of interest. Based on individual preferences, generators of interest then decide to devote their interest to the given subject. On the other hand, the subjects of interest generate events, which have the potential to attract interest from the generators of interest. Those events either manage to attract interest or stay unnoticed. Therefore, interest as a result of an event, cannot be assumed to produce negative values. The bottom line must be considered as zero addition interest — a full absence of interest for an event. This point of full absence of interest might be reached in the following three cases.

In first case, the given subject of interest does not generate events, which attract interest. There are no outgoing events, which would have as consequence an attraction of interest by the generators of interest. In the second case, however, the given subject of interest generates events, which have the potential to attract interest, but for some reason fail to do so. The third case implies events, not related to the subject of interest, which distract interest from the given subject of interest. This would imply, that another subject of interest attracts a large amount of interest, such that the generators of interest focus their interest to this other event, leaving the given subject of interest without any interest — zero interest.

Now consider the opposite case — the presence of interest. The presence of interest might be cause by the following three cases. In the first case, the subject of interest generates some event, which lead to an attraction of interest by the generators of interest. The second case implies an external event, which leads to interest devotion by the generators of interest. In order capture and measure this case GT includes in related queries. In the third case, the subject of interest has managed to attract interest in the past. This has led to a constant process of devotion of interest by the generators of interest. The existence of such a process among the given TBNVs has to be assumed due to the fact that they have already attracted interest by making it into the AI 100 listing. This process of constant interest by generators of interest might be thought of as the underlying data generating process in the time series.

The discussion of presence or absence of interest leads to the conclusion, that the only way for a subject of interest to generate interest by itself, is to generate events, which then eventually attract interest. However, one should bear in mind, that the second case of interest generation implies, that there is some ongoing interest generation process, into which those additional events fall. This ongoing interest generation process might be a series of zero values or a constant fluctuating process around some stable value. The question arises of how to distinguish the attracted interest of an additional event from the general interest generation process, which is assumed to be present.

Visually it has been possible to identify correlation between funding events and GT interest. The question arises of how to establish this correlation mathematically. Furthermore, it will be necessary to establish statistical evidence of classify those outliers as indeed shocks, which not follow the underlying data generating process. Establishing a measurement for correlation of outliers, caused by funding events, allows for a general assessment of correlation. At the same time this might be used as a tool to compare the responsiveness of time series data to funding events across time series.

Earlier the question was raised of how to assess the "quality" of time series data. In the framework of this thesis, it was decided to use responsiveness of time series data to funding events as a measure for quality. Note, this is in-line with the description of alternative data. Alternative data is suppose to mimic an ongoing process inside the subject of investment - in this case inside the new venture. Therefore, the more the time series data of interest mimics the investment process of venture capital funding, the better web-search traffic qualifies as alternative that. On might not only focus on overall responsiveness of time series data, but also on differences in responsiveness among different funding events.

This measuring tool was developed based on the following assumption. After deleting the seasonal component and known outliers the time series was described by an ARIMA model. The resulting residuals are assumed to be normally distributed. Due to normality, it becomes possible to compare the distribution of the period before a given funding event, including this funding event, with the value of that funding event itself. Then it is calculated how likely it is that indeed the value of the funding event was drawn from the normal distribution before the funding event. This is done by calculating the z-score at the funding event. The z-score or standard score is the distance from the mean, measured in standard deviations, assigned to a value, assuming that this value was drawn from a normal distribution with given a mean and variance. Note, that this measure allows comparing possible outbreaks at different points in time for one time series and across different time series.

Find below the mathematical definition of the standard score and how it is applied in this thesis. To visually translate the mathematical concept, consider figure 15. Shown are 3 time periods of type-II residuals, resulting from the ARIMA process as described in the previous sub-chapter. To remind, these residuals are obtained by taking the time series data with event shocks, but where the seasonal anomaly was eliminated. Then subtracting the prediction from the ARIMA process, which was identified as best fitting. Therefore, the residuals in figure 16 display large outliers, highlighted in red, which were caused by a funding event, taking place at the time point in time. These outliers will be subject to the measuring process, which is described in the following.

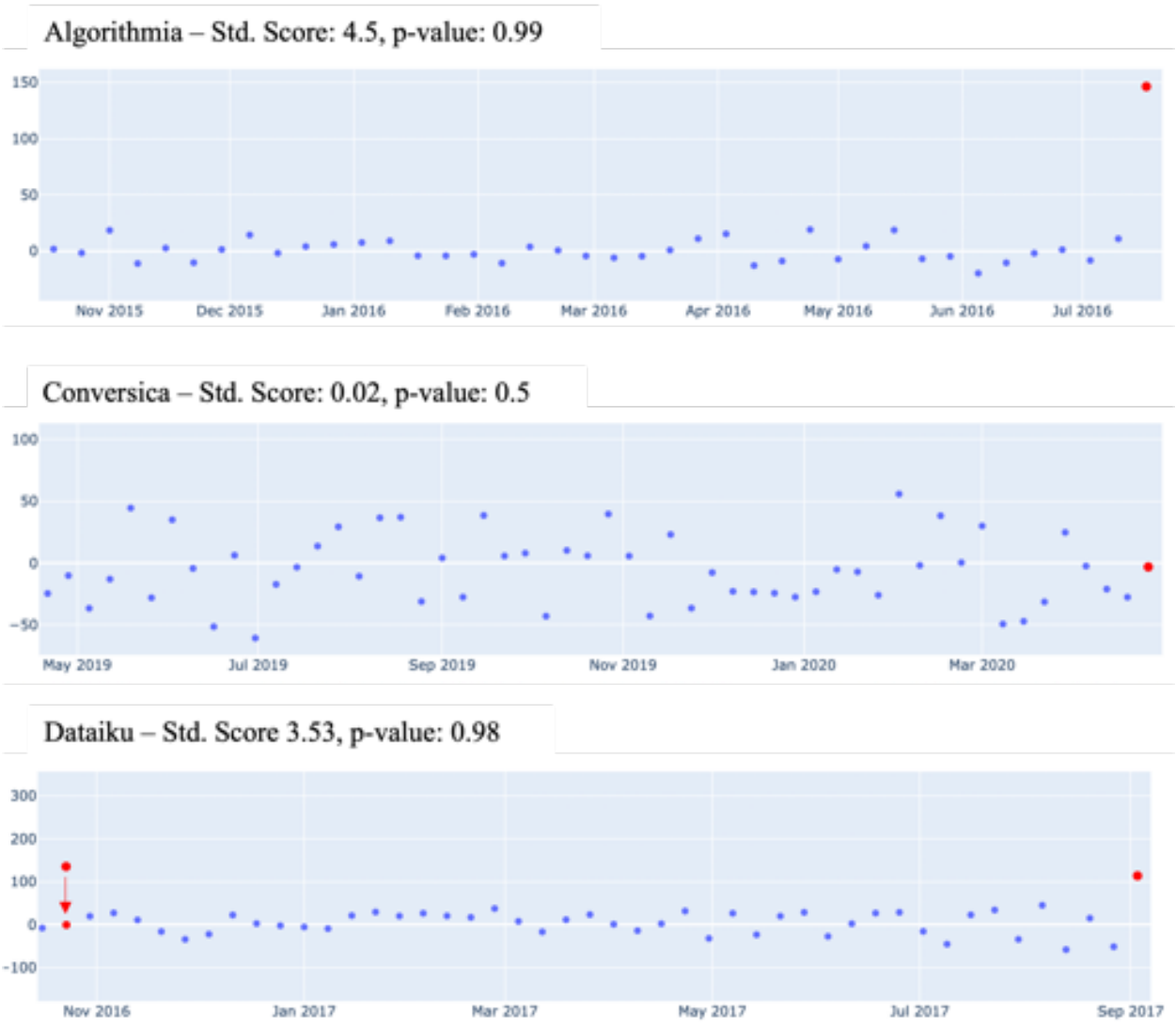

Figure 15: Time Periods of Residuals Including Shocks by Events

Consider a given outbreak in a given period. Let the time index of this outbreak be t. Then the value of the data point at this outbreak is x_t. For determining the standard score of x_t, an interval is required, where $|\mathbf{x}| = n + 1$, where n is the length of the interval before the event values, x_t.The values x_{t-n}, x_{t-n-1}, ..., x_{t-1}, x_t are assumed to be $iid\, N(\mu, \sigma^2)$. As μ and σ are unknown, the sample mean $\bar{X}_t$ and sample standard deviation s_t will be used for approximation. The length of the event interval, $n + 1$, is crucial for the calculation of $\bar{X}_t$ and s_t. Due to the properties of both sample statistics, outliers in small samples have an inflating effect on their value. On the other hand, outliers in large samples have a relatively small effect on the value of the sample statistics. As x_t is assumed to be an outlier, it is necessary to pick n relatively large to avoid an inflation of $\bar{X}_t$ and s_t. However, n should be logically embedded in the larger framework of the thesis. The length of the event interval, n, was chosen based on an analysis of all interval lengths between the funding events of the 115 companies. For all companies, the length of time intervals between their funding events was calculated in weeks. The distribution of the resulting lengths of all intervals is illustrated in figure 16. Note, the distribution is not centered, but rather congested towards 0. Therefore, it was decided to pick $n = 37$, which is the median of the resulting distribution. Therefore, the measure of the outbreak, caused by funding events, is defined as the standard score. The standard score, z_t, is defined as the standardization of x_t under the assumption that $x_t \sim N(\mu, \sigma^2)$.

$$\mathbf{x} = x_{t-n},\, x_{t-n+1},\, \ldots,\, x_{t-1},\, x_t \sim iid\, N(\mu, \sigma^2),\ n = 37$$

$$\bar{X}_t = \frac{\sum_{t-n}^{t} x_i}{n+1} \to \mu, \quad s_t = \sqrt{\frac{\sum_{t-n}^{t}(x_i - \bar{X}_t)^2}{n}} \to \sigma$$

$$z_t = \frac{x_t - \bar{X}_t}{s_t}$$

In figure 15, the first event interval relates to a funding event of the new venture "Algorithmia". Displayed is rather stable data generating process with a large outbreak in $t = third\, week\, of\, July$ 2016. By the definition of the standard score this results in $z_t = 4.5$. Under the assumption of $x_t \sim N(\mu, \sigma^2)$, the standard score z_t might be also interpreted as z-statistic, under the under the null hypothesis $H_0 : x_t = \mu$. For $z_t = 4.5$ the corresponding p-value is 0.999. Therefore, the null hypothesis of $x_t \sim N(\mu, \sigma^2)$ is rejected at a significance of 0.001%. Consequently, the data generating process $N(\mu, \sigma^2)$ has clearly been interrupted in the third week of July 2016. The data point in the third week of July 2016 does not follow the assumed normal distribution in $\mathbf{x}$.

	Obs.	Std. Score	Med. p	Sig. 5%	Sig. 10%	Sig. 15%
Investment Series						
Seed	22	1.54	0.87	32%	45%	50%
Seed VC	65	2.03	0.95	51%	58%	65%
Seed VC - II	16	1.38	0.81	25%	25%	44%
Series A	111	2.36	0.99	57%	65%	71%
Series A - II	27	1.81	0.97	56%	59%	63%
Series A - III	10	1.40	0.94	50%	60%	60%
Series B	99	2.47	0.99	67%	74%	79%
Series B - II	19	1.75	0.94	42%	58%	58%
Series C	71	2.84	0.99	72%	77%	86%
Series C - II	21	1.77	0.95	48%	67%	71%
Series D	41	2.14	0.98	63%	71%	73%
Series E	20	2.22	0.99	60%	75%	85%
Exit Events						
Acquired	10	4.24	1.00	100%	100%	100%
IPO	10	5.62	1.00	100%	100%	100%
Other Events						
Unattributed VC	12	0.72	0.73	25%	33%	33%
Loan	14	0.75	0.77	7%	7%	29%
Grant	10	1.14	0.76	40%	40%	40%
Incubator/Accelerator	31	0.80	0.70	29%	32%	39%
Incubator/Accelerator - II	17	1.07	0.86	18%	35%	59%
Aggregation						
All Events	760	1.99	0.96	51%	59%	65%

Table 1: Metrics by Event Groups

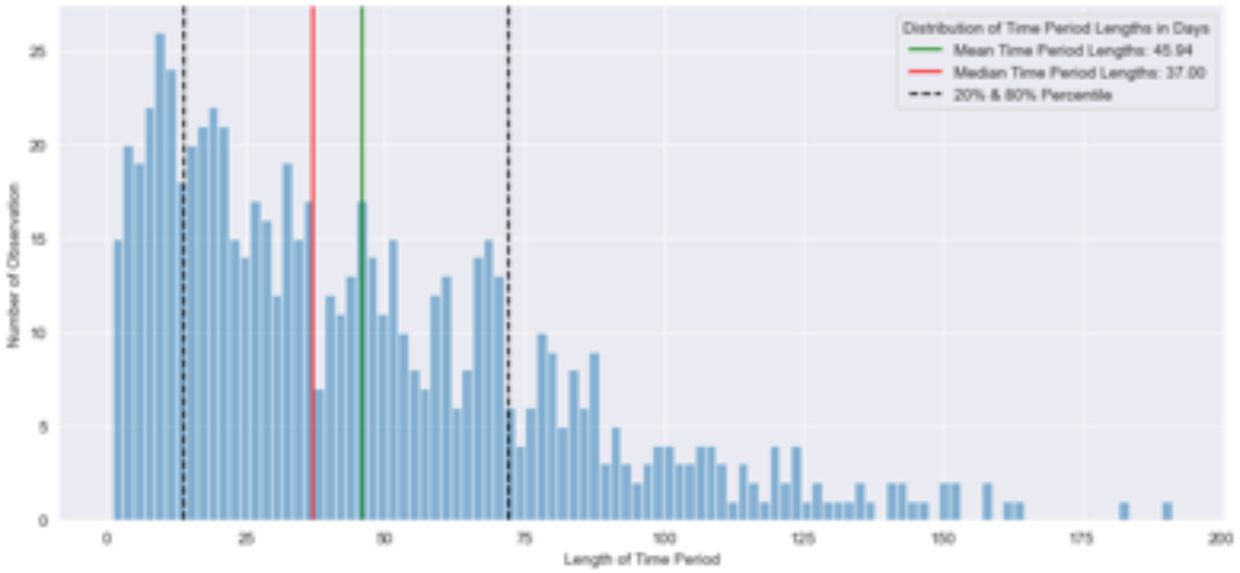

Figure 16: Distribution of Time Intervals Between Events

The second event interval in figure 16 displays a more volatile data generating process. The event value, x_t, at $t = fourth\ week\ of\ April$ 2020 is not visually distinguishable from earlier values. By the definition of the standard score this results in $z_t = 0.02$. Under the assumption of $x_t \sim N(\mu, \sigma^2)$ took an reasonable value. The corresponding p-value is 0.5. Therefore, the null hypothesis of $H_0 : x_t = \mu.$ is not rejected.

The final event interval, $\mathbf{x}$, displays a similar dynamic as in the first event interval. However, note that into the event interval falls another data point, which is related to another funding event. As already mentioned, this might lead to unnecessary inflation of the sample statistics. Since the length of all event intervals was chosen to be 37, there are cases where the true event interval from x_t to the previous funding event $x_{t^{previous}}$ is less than 37. Knowing that the other data point is related to another funding event, it is assumed to be an outlier and, therefore, averaged, as the red arrow illustrates. This approach was applied in all such cases.

For the 115 companies, 760 funding events were recorded. The Standard score, with respect to the 37 foregone weeks, was calculated. Figure 17 shows the resulting distribution of standard scores and associated p-values of all 760 funding events. The distribution of the standard scores is centered around 1.99. Therefore, for further inference, the mean of this distribution will be consulted. The distribution of the p-values congested towards 0. Therefore, for further inference, the median will be consulted. Table 1 shows the resulting

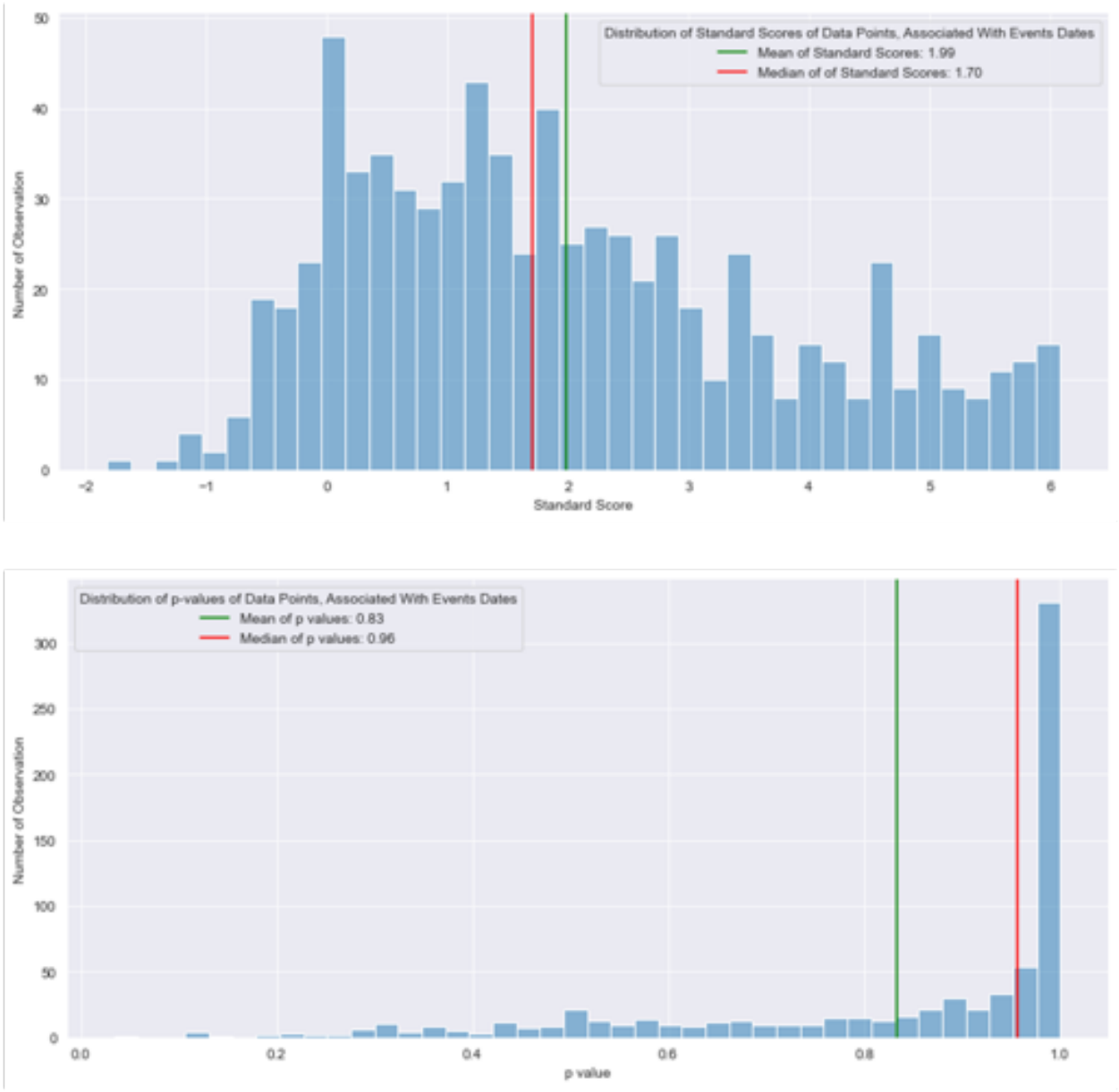

Figure 17: Distribution of Standard Scores and p values of Event Data Points

statistics grouped by different funding events. The funding event dataset obtained from CB insights contained a large number of rather undefined funding events. In table 1, events are displayed, that were observed more than 10 times.

5.4 Structural Changes After Funding Events

The previous subsection has made it undoubtedly clear that funding events have a measurable effect on web-search traffic related to new ventures, which is represented by GT data. However, it remains unclear what kind of long-term influence the funding events cause. This sub-section concerns the question of the presence of structural changes in the time series data.

A common approach form measuring the long-term effect of all kinds of events in time series data is the analysis of structural change. The events suspected to have long-term influence are also called breaks. These breaks could cause a change in the mean or in other parameters of the time series. In economics, a structural break might occur when a war breaks out, a major change in government policy is invoked, or some equally sudden and radical event. In the context of this thesis structural breaks in the time-series data might be interpreted as an event that leads to a significant rise in the mean of interest in a new venture. This would imply that some events have the potential to attract a significantly larger amount of long-term interest. Another possible long-term effect might be a significant long-term change in the growth of interest. In terms of time series analysis, terms like intercept and slope are used to describe the mean and growth rate, respectively.

Figure 18 lists the four possible outcomes of a structural change, where it is assumed that a time series might be described by an intercept and slope. To correctly understand figure 19, in terms of structural changes, note that both regressions displayed are not bounded to time. They are displayed on the same x-axis, however, they describe two processes that happen independently from each other in a sequence after a breakpoint. For further illustration refer to the regression, marked by the long dotted line as the first in the sequence, while the regression marked with the dotted line as the second in the sequence. Outcome A displays the case of coincident regressions. This case describes the process with no structural changes.

The coefficients for the intercept and slope of both regressions are the same. It is concluded that the suspected breakpoint did not cause any change to the data generation process of the time series. Outcome B displays the case of parallel regressions. A change in the intercept between the regressions is observed. This implies the breakpoint caused a change

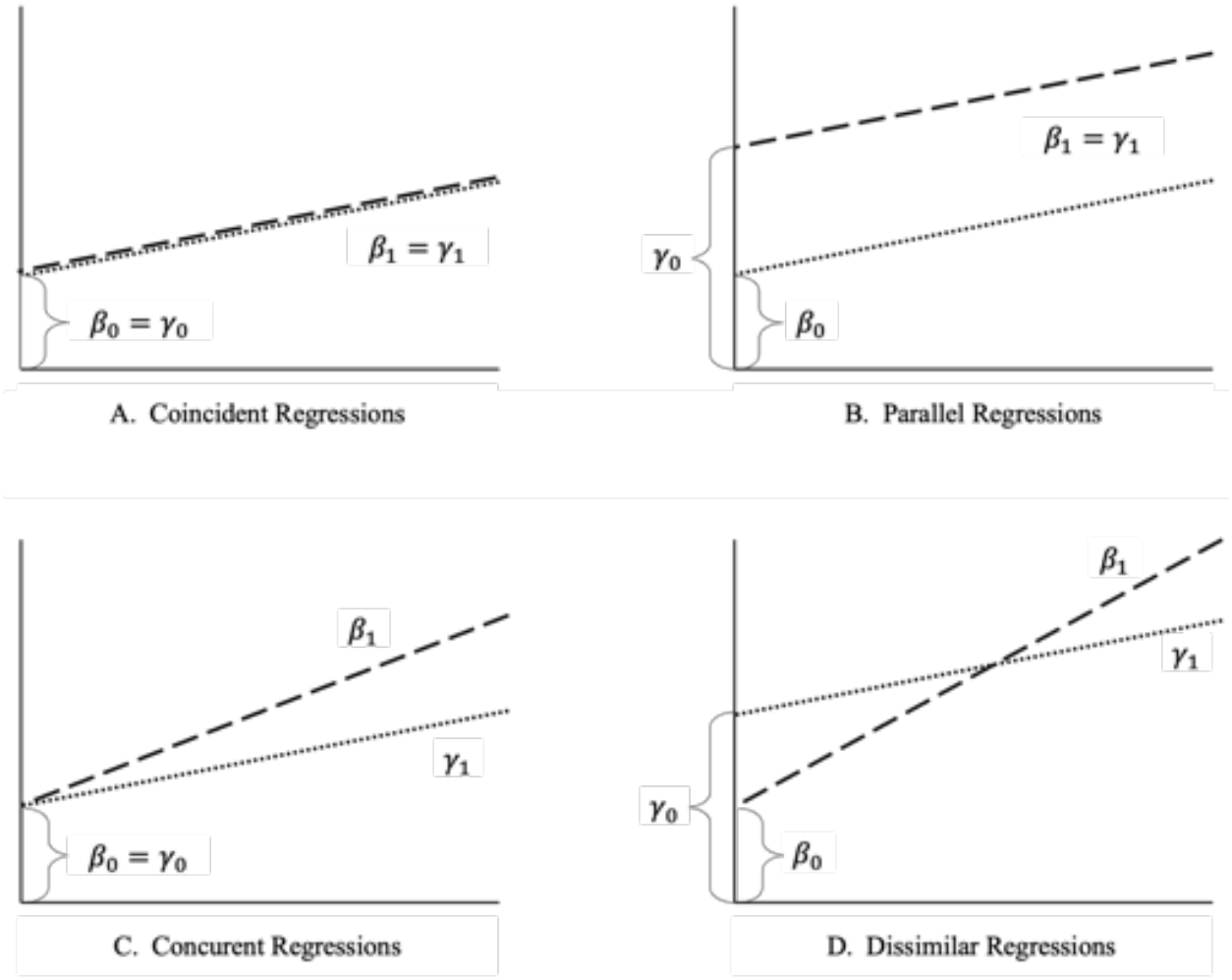

Figure 18: Possible Outcomes of Dummy Regression

in the mean of the data generation process of the time series. However, note that the slope of both coefficients remains equal, implying that the growth rate in the process stays unchanged. Outcome C displays the case of concurrent regressions. This implies that after the breakpoint the growth rate in the data generating process changed, while the intercept remains unchanged. Outcome D displays the case of dissimilar regressions. This implies that both changes observed in B and C take place at the same time. The position of breaks in time series data might be known or unknown. In this thesis, it is assumed that all possible breakpoints are known. Possible breaks are represented by the time points of the observed funding events. A possible breakpoint will be classified as a true breakpoint if either outcome B, C, or D is true. However, it might well be the case that other breakpoints are present in the time series data, which, however, are caused by other events besides the funding events. Indeed, in the following chapter of this thesis, such events will be discussed. Moreover, it is assumed that the variance in the time series data remains stable after the suspected breakpoints.

To conduct the analysis of structural changes time intervals need to be defined on the left and right-hand side of the breakpoint. In figure 16 the distribution of the time intervals between funding events was plotted. The black dotted lines mark the boundaries of the interval containing 60% of all values. The corresponding values are 14 and 72, respectively. It was decided to allow only for intervals, whose length is located over the 20% percentile. Periods, whose length is located over the 80% percentile, were trimmed to the border of the 80% percentile. There are two time periods, τ_1 and τ_1, which form together the full time period, τ. Then length of the full time period is $n = n_1 + n_2$.

$$\tau \;=\; \tau_1 + \tau_2, \;\; \textit{where} \;\; |\tau_1| = n_1 \;\; \textit{and} \;\; |\tau_2| = n_2$$

$$\tau_1 \;=\; [t - n_1, \ldots, t-1]\,, \;\; \tau_2 \;=\; [t,\; t+1,\; \ldots,\; t+n_2]$$

To build up the mathematical foundation to statistically establish whether one of the three cases holds true for a given possible breakpoint, a dummy regression was constructed. The endogenous variables are x_0 and x_1, where x_0 represents the intercept and x_1 the slope of the dummy regression. The length of the examined period, n, is defined by the lengths of both sub-periods, n_1 and n_2.

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \; \mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \\ \vdots \\ n \end{bmatrix}, \; \Delta = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

$$where\ \Delta_i = 1\ if\ i \in \tau_2$$

To capture the suspected structural changes in the time series after a possible breakpoint, it is necessary to build duplicate the regression process. As described in figure 18, two sequential regression processes will be constructed. The breakpoint in time is given by t. Therefore, a dummy variable was created, Δ, with a value equal to 1 after the funding event at time, $i = t$. This dummy variable is then used for duplication in the regression. Due to the duplication, two new parameters are created, γ_0 and γ_1, which measure possible changes in the intercept and slope, respectively, starting from $i = t$. The final regression equation looks as follows.

$$y\ =\ \beta_0 \mathbf{x}_0 + \beta_1 \mathbf{x}_1\ +\ \gamma_0 \mathbf{x}_0 \Delta + \gamma_1 \mathbf{x}_1 \Delta\ +\ \epsilon$$

To determine the outcome of the regression, as shown in figure 18, the p values of the t-test from the duplicated parameters, γ_0, γ_1, were consulted. Recapitulate the null-hypothesis of the t-test, $H_0 : \gamma = 0$. Therefore, at a given significance level, α, the possible outcome in figure 19 are:

$$\begin{array}{l} p\,value_{\gamma_0} > \alpha,\ p\,value_{\gamma_1} > \alpha \rightarrow Coincident \\ p\,value_{\gamma_0} <= \alpha,\ p\,value_{\gamma_1} > \alpha \rightarrow Parallel \\ p\,value_{\gamma_0} > \alpha,\ p\,value_{\gamma_1} <= \alpha \rightarrow Concurrent \\ p\,value_{\gamma_0} <= \alpha,\ p\,value_{\gamma_1} <= \alpha \rightarrow Dissimilar \end{array}$$

The dummy regression was performed on all funding events, which allows for time periods of minimum 14 data points on both sides. It was possible to run the regression for 555 funding events. it was decided to divide the 555 funding events into two groups. In the forgone analysis in the previous sub-chapter, the deviation caused by the funding events from the ground-lying data generation process was measured. With the help of p-values, it was analyzed which events display a significant deviation from the ground-lying data generation process. Therefore, one group contains all events, which display a significant deviation at $\alpha = 5\%$. This group contains 294 observations. The other group contains all events, which do not display a significant deviation at $\alpha = 5\%$. This group contains 261 observations. The results of the structural analysis are summarized in table 2.

	Significant Events			Insignificant Events		
Outcome	Sig. 1%	Sig. 5%	Sig. 10%	Sig. 1%	Sig. 5%	Sig. 10%
Coincident	231	167	133	225	185	162
Parallel	13	34	45	10	26	27
Concurent	43	65	77	22	39	37
Dissimilar	6	27	38	1	8	32

Table 2: Resulting Outcomes of Structural Breaks by Significance

5.5 Changes in Variance after Funding Events

Another common long-term effect in time series data, after a given breakpoint, is a change in the variance of the underlying data generating process. In time series analysis, this phenomenon might be interpreted as the presence of heteroscedasticity. Heteroscedasticity is described as a non-constant standard deviation of a predicted variable over time. Heteroscedasticity mostly arises in two forms, namely conditional and unconditional. Conditional heteroscedasticity describes non-constant volatility in relation to the prior period's volatility. For example, when a steady increase in volatility from week to week is observed. On the other hand, unconditional heteroscedasticity refers to general structural changes in volatility that are not related to prior period volatility, but rather to a fixed point in time. Unconditional heteroscedasticity is present when separated periods of significantly different volatilities can be identified. In terms of the current analysis, there are clearly identified possible time funding events, which might be identified as a breakpoint in the volatility of the data generating process. This sub-section concerns possible changes in the variance of the underlying data generating process caused by funding events.

Measuring the outbreaks related to funding events, however, has shown that not all funding events have a related significant outbreak. This raises the question of whether an unnoticed funding event can cause a change in variance. This question will be subject to further discussion in the next chapter. In the current analysis, significant outbreaks are analyzed separated from insignificant outbreaks. In the second sub-section of this chapter, 115 time-series were analyzed and decomposed. The result was cleaned time series data. On this cleaned data the best-fitting ARIMA composition was determined. The data generating process of the time series data was described by best-fitting ARIMA processes. Type-I residuals will be used in this part of the analysis.

Two time intervals were defined for the analysis. One interval before the funding event, τ_1, and the other after , τ_2. In terms of the following variance test, those intervals will be referred to as samples. The funding event takes place at time, t. The lengths of both intervals were chosen by the same conditions as for the structural changes. The sample standard deviation will serve as an approximation for the true variance of the intervals. That is,

$$\tau_1 = [t - |\tau_1|, \ldots, t-1]\,,\ \tau_2 = [t, \ldots, t + |\tau_2|],\ |\tau_1|,\ |\tau_2| \ \in \ \{14, 15, \ldots, 72\}$$

$$s_1^2 = \frac{\sum_{t-|\tau_1|}^{t-1}(x_i \ - \ \bar{X}_1)^2}{|\tau_1|} \ \rightarrow \ \sigma_1^2$$

$$s_2^2 = \frac{\sum_{t}^{t+|\tau_2|}(x_i \ - \ \bar{X}_2)^2}{|\tau_2|} \ \rightarrow \ \sigma_1^2$$

It was decided to conduct different variance test on the data - Levene's test, the Bartlett's test and the F-test. Levene's test is used to test if several samples have the same variance. The Levene test is robust to deviations from normality in the samples. This is a two-tailed test. Thus, it is not possible to test in directions. There are only samples. Therefore,

$$(1)\ H_0 : \sigma_1^2 = \sigma_2^2,\ \ H_1 : \sigma_1^2 \neq \sigma_2^2$$

Bartlett's test is used to test if several samples have the same variance. Bartlett's test is sensitive to departures from normality. This will be taken into account at when comparing the the results from Levene's and Bartlett's test. This is a two-tailed test. Thus, it is not possible to test in directions. In the given case there are only two samples. Therefore,

$$(2)\ H_0 : \sigma_1^2 = \sigma_2^2,\ \ H_1 : \sigma_i^2 \neq \sigma_j^2$$

The F-test is used to test if the variances of two populations are equal. This test can be a two-tailed test or a one-tailed test. The two-tailed version tests against the alternative that the variances are not equal. The one-tailed version only tests in one direction. Therefore,

$$(3)\ H_0 :\ \sigma_1^2 = \sigma_2^2,\ \ H_1 :\ \sigma_1^2 < \sigma_2^2$$

$$(4)\ H_0 :\ \sigma_1^2 = \sigma_2^2,\ \ H_1 :\ \sigma_1^2 > \sigma_2^2$$

The choice is determined by the problem. In the given framework, a change in variance after an funding event is suspected. However, it is of interest whether the possible changes result in a period with larger or lower variance. Therefore, both versions of the one-tailed test will be conducted.

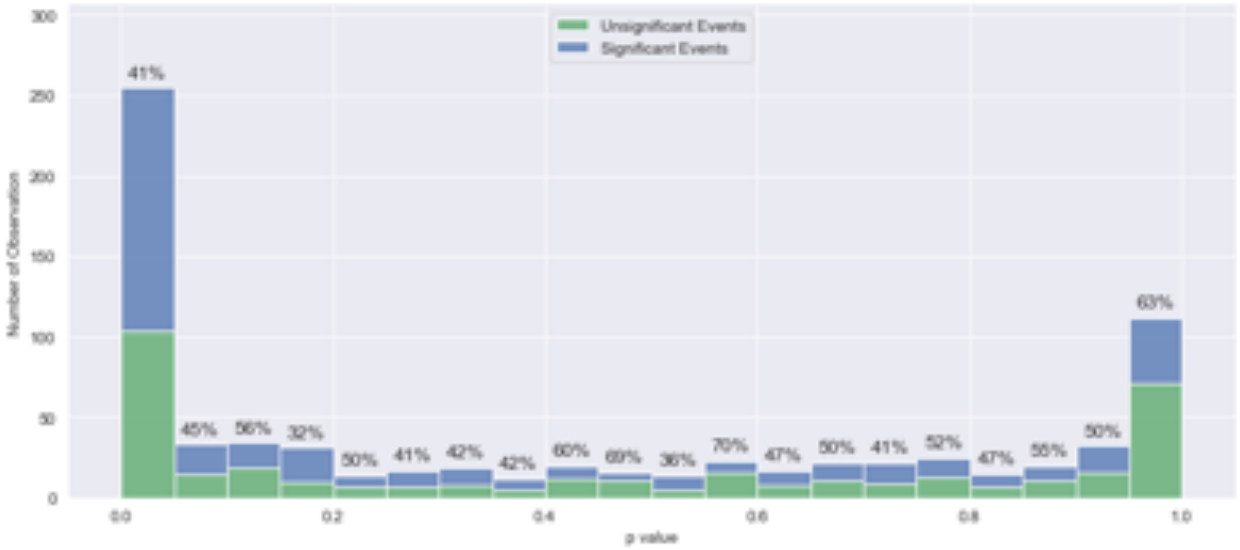

Figure 19: Distribution of p-values of F-Tests by Significance of Events

The four tests were conducted on the ARIMA residuals for the 555 funding events, which displayed high significance in the outbreak. The resulting distributions of the p values are shown in figure 19. Not that figure 19 shows the distribution of the p-values by two groups. The first group contains the p-values of events, which have been classified as significant outliers. The second group contains the p-value of events, which have not been classified as outliers. On top of the bars the percentage of the insignificant group, compared to all events in this bar is displayed. The overall results are summarized in table 3. It was decided to group the resulting rejection rates of the null hypothesis by event groups. The first group displays the sequence of the event. That is, the order, in which a given event took place. The second group displays the funding round an event is related to.

		Levene $H_1: \sigma_1^2 \neq \sigma_2^2$		**Bartlett** $H_1: \sigma_1^2 \neq \sigma_2^2$		**F-test** $H_1: \sigma_1^2 < \sigma_2^2$		**F-test** $H_1: \sigma_1^2 > \sigma_2^2$	
Event	**Obs.**	**5%**	**10%**	**5%**	**10%**	**5%**	**10%**	**5%**	**10%**
Event 1	55	53%	56%	53%	58%	58%	62%	0%	2%
Event 2	54	31%	37%	41%	52%	46%	50%	6%	7%
Event 3	59	20%	32%	34%	41%	27%	36%	10%	12%
Event 4	54	13%	15%	24%	30%	19%	26%	9%	9%
Event 5	46	9%	20%	15%	20%	13%	22%	9%	9%
Event 6	37	8%	8%	24%	32%	19%	24%	14%	16%
Event 7	30	10%	10%	13%	20%	17%	27%	3%	3%
Event 8	22	18%	23%	14%	14%	14%	18%	0%	5%
Seed VC	33	61%	67%	64%	73%	70%	76%	0%	3%
Series A	63	48%	56%	62%	65%	59%	60%	6%	11%
Series B	66	23%	33%	32%	41%	27%	36%	14%	15%
Series C	51	20%	33%	35%	39%	25%	29%	14%	20%
Series D	26	4%	19%	31%	35%	19%	27%	15%	15%

Table 3: Rejection Rates of Change in Variance by Event Groups

6 Discussion

The previous chapters provided the theoretical and analytical framework of this thesis. Investment activity and its current challenges in the venture capital industry were introduced. The necessity for alternative data in the venture capital investment process was formulated. Based on previous research, Google Trend was suggested as possible alternative data, for counteracting the scarcity of data in the venture capital industry. Finally, different kinds of statistical tests were conducted to explore the influence of investment activity on GT data. This chapter yields to interpret and compare the statistical result of the analytical part of this thesis. Those results are then discussed and their implications are related to the theory of venture capital. Firstly, anomalies in the data, like outbreaks and uncommon patterns in the data are discussed. Secondly, the magnitude caused by funding events is inspected. Finally, long-term influences are analyzed and put into context with the topic of this thesis.

6.1 Suspicious Looking Time Series

The visual investigation of the time series in figure 10 and 11 of chapter 5.1 has shown the suspicious appearance of some time series. Consider the time series from "ALICE Technologies" and "Atomwise". There are several details in this graph, based on which this data might not be classified as a time-series data. However, it is non-trivial to define a coherent method by which to classify certain time series as "not a time series". This discussion becomes more complicated when considering the responsiveness to funding events of "Atomwise". To claim the "uselessness" of this type of time series data might not be appropriated. The analysis of outbreaks has proven that even this type of time-series data can display high responsiveness to related funding events. This stands in contradiction with Malyy et al. (2021), whose algorithm for time series data selection purposeful discriminates for such data. The main problem with this type of data in the framework of this thesis arises during the regression process. The ARIMA processes, which were identified to be optimal had orders of (0,1,1) and (4,1,4), respectively. However, is not obvious how this corresponds to the graph. I assume that both time-series follow a (0,0,0) ARIMA process. Therefore, the raw time series data is the underlying data generating process. Indeed, no AR nor MA component can be visually observed. Moreover, the t-statistic of the corresponding estimated coefficients would likely imply insignificance due to the large number of zero values in the data. A different type of regression model might better describe the true data generating process. However, for the goals of this thesis, this approach is sufficient. The predicted time series sufficiently display the dynamics of the data generating process to compare it with outbreaks.

6.2 Seasonal Anomalies in The Data

The visual investigation of the time series in figure 9 in chapter 5.1 showed a strong decline in interest for new ventures at the end of every year. Further analysis with the help of run-sequence plots in figure 11, revealed that the given anomaly cannot be classified as a seasonal component. Therefore, it was decided to eliminate this anomaly from the data. However, the question arises of how this anomaly might be classified and what the consequences are for this thesis.

Figure 12 showed that this anomaly is not only to be observed among the search queries, which are used in this thesis. Indeed, all kinds of search queries display this anomaly. Therefore, the causation of this anomaly is to be found outside the framework of this thesis. Considering that 85% of the companies subject to this thesis are American, it might be assumed that the anomaly is caused by the absence of interest for the given search terms around Christmas and new years eve. It is not of interest to this thesis to analyze the question of whether there exist search terms, which do not display this anomaly or even display a positive dynamic around that time.

As shown in figure 11, the end of every year displays always a local minimum in the data. Even though it is possible that those local minima are not sufficient statistically significant to classify them as outliers. It is known, however, what causes them and as the causation is due to an outside event, it is legit to eliminate this anomaly.

Nevertheless, as mentioned in chapter 4.3.1 the time series data obtained from GT represents the worldwide interest in a search term. Note, that 15% of the companies, however, are non-American and might therefore display anomalies linked to the country, in which they operate. There are methods to analyze the share in the interest of some world regions, compared to others. This would allow filtering more precisely such possible anomalies. However, this is beyond the scope of this thesis. This is an uncertainty factor in the analysis of this thesis, which needs to be accounted for.

The description of the term alternative data mentions that the data generating process in the alternative data should mimic some process in the company. However, the absence of interest due to the Christmas holidays and new years eve should not play a role in the VC investment process. Therefore, it is concluded that a correction of this anomaly is necessary for any analysis of GT data for the VC investment process.

6.3 Unnoticed Outliers

The analysis in this thesis covered the implication of known outliers caused by funding events. However, it needs to be assumed that there are more outliers, which might not be related to funding events. This raises the question of outlier identification, which is a field of research by itself. A simple visual inspection, however, can also be performed to visually identify other outliers. However, due to a small sample size of 55 events, no further analysis was conducted.

Such an analysis was conducted for a subset of companies to address this question. It was possible to identify 55 media events linked to the visual outbreaks in the data. The median p-value of those events is 0.999. However, this value is biased as it was constructed using visually significant outliers. It might very well be the case that other media events have stayed unnoticed. Indeed, the theoretical problem arises of how to solely mathematically classify outliers, without can further background knowledge. As mentioned before, this is a field of research by itself. However, to conduct such an analysis, I propose deleting the seasonal anomaly and all other known outliers. Then detrending the time series data. Then every data point can be analyzed sequentially on the resulting time series. Assuming the data is normally distributed the z-score method can be applied to all data points sequentially. By considering the resulting p-values for every data point in the time series, outliers might be identified. Then for every significant data point, a Google search has to be conducted. The goal is to find a media event that happened at the point in time of every significant outbreak. If a media event is found, that media event can be taken as causation for the outbreak in interest.

6.4 Magnitude in Interest of Funding Events

The analysis of the magnitude of outbreaks in the time series data caused by related funding events has undoubtedly proved the link between web-search traffic and investment activity in the venture capital industry. The median of the p-values of all events is 0.96. The median is the value separating the higher half from the lower half of a probability distribution. Therefore, at a significance of $\alpha = 4\%$ it can be assumed that 50% of all funding events cause an outbreak in interest. This implies that for new ventures investment activity plays a major role in interest in a company. This might have been expected. However, it was firstly analytically and statistically proven in this thesis.

The discussion in chapter 5.3 of subjects of interest and generators of interest needs to be continued. As mentioned, generators of interest devote their interest due to individual preferences. In the given case on might extend the discussion to characterize the generators of interest further. One can assume two major groups in the context of new ventures. Customers of the new venture and the VC community. The interest of both groups does not coincide. Customers do not participate in the investment process and investors do not buy the offered products by the new venture. Therefore, it must be assumed that the deviations from the outbreaks during investment events are exclusively caused by the VC community. This implies that the outbreaks exclusively display the VC investment process.

Indeed, consider the standard score of the outbreaks. In table 1 an increasing average standard score from Pre-Seed to Seed to Series A, B, and C is clearly observed. This trend becomes negative when moving to further series D and E. This is clearly in accordance with the theory of funding events discussed in chapter 2.1. This is one more outstanding finding. This suggested that the interest recorded during outbreaks theoretically displays the dynamics of the valuation of new ventures. In addition to Malyy et al. (2021), this finding suggests a correlation between valuation points and outbreaks in interest during investment events.

Recall figure 8. Not at every funding event, a valuation of a company is provided. To access the valuation of a new venture during a funding event that is not accompanied by a valuation, the outbreak in interest in GT data might be consulted. For further research, I suggest comparing available funding points with corresponding outbreaks in interest. This might provide an alternative valuation of the new venture for the case when no valuation was conducted.

Now refer back to the question of qualification of visually suspicious time series data. Recall figure 17. Consider the distribution of the standard score. One can observe a large number of events, which have caused a negative standard score in the deviation from the data generating process. This is logically not possible. Any event, even a negative one can only create an increase in interest. Therefore, the question remains of how to classify this type of event, causing negative standard values. I propose to classify them as completely ignored. They were completely ignored and by coincidence fell in a week where the interest was at a local minimum. Note, this implies that there is an unknown number of events, which stay completely ignored and randomly fall into some weeks. It can be assumed that the same amount of completely ignored events fell into weeks where the interest was on a local maximum.

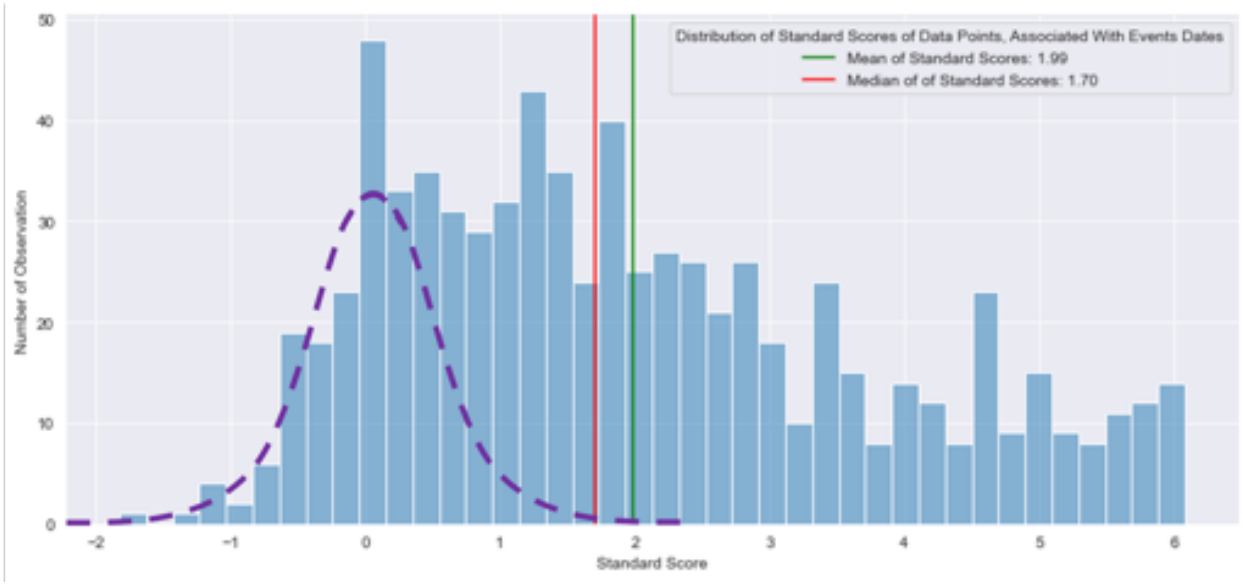

Figure 20: Assumed Distribution of Totally Ignored Funding Events

As it is assumed that the underlying data generating process is normally distributed such an ignored funding event can fall with equal probability into a week with a value below or above the mean of the data generating process. The emphasis is on "with equal probability". Therefore, on average a certain amount of ignored events with a negative standard score is observed. On the other hand, there is a proportional comparable amount of ignored events with a positive standard score. However, this amount is not observed. Only the left-hand side of this distribution is observed. Due to the assumption of normality, it is possible to guess the right-hand side of this distribution. The assumed distribution is displayed in figure 20.

This discussion is of further interest, as the events falling into left hand side of the distribution might reveal further insights, when trying to understand what characterizes ignored events. After a further analysis into this topic, it was found that events, falling on the left hand side of the violet distribution have in common, that the new venture, they are associated to, display low responsiveness in all funding events. For every new venture the average p-value of events was calculated. There are 22 new venture, which display an average p-value of lower than 0.75. Among those 22 new ventures, 95% of them have events with negative standard scores. To compare, from new ventures, with an average p-value of higher or equal to 0.75, only 37% of them have events with negative z scores. This leads to the assumption, that the given subgroup of 22 companies, might be considered as low-responsive to non-responsive

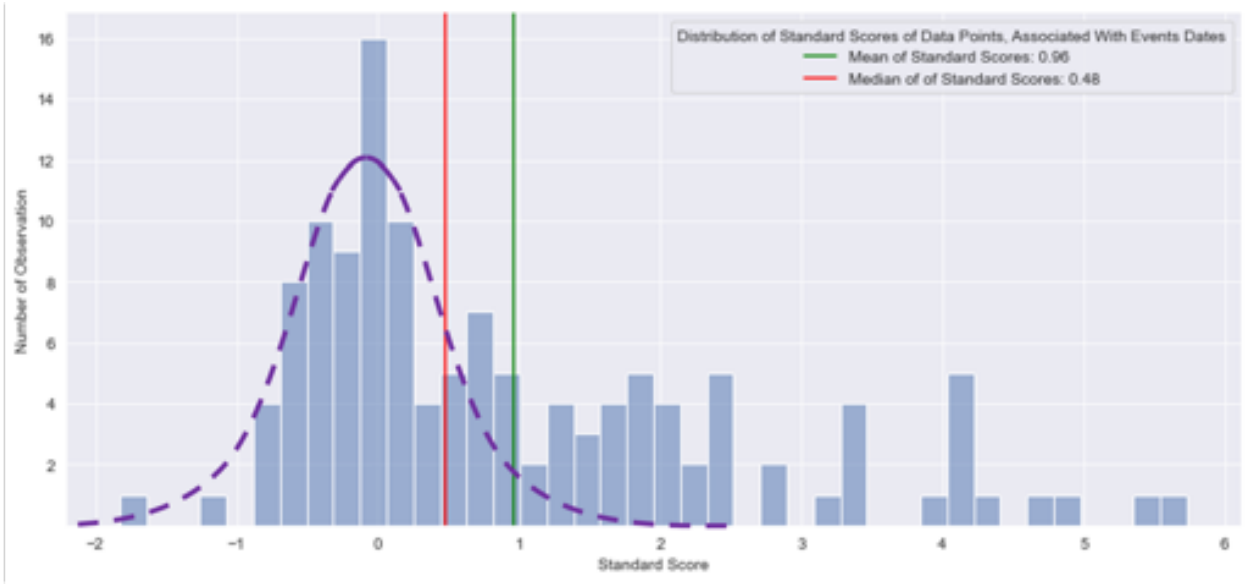

Figure 21: Standard Score Distribution of Low-Responsive New Ventures

with respect to investment activity. To analyse this topic further, consider figure 21. Figure 21 shows the distribution of the standard scores from low-responsive new ventures. One can observe that the distribution is significantly closer to normal, compared to the overall distribution of standard scores. Bearing in mind that ignored events can fall on any value of the normally distributed underlying data generating process, this has as implication, that low-responsive new ventures might be analysed separately, or even excluded from analysis. However, there are some events even among low-responsive new ventures with significantly high standard scores.

6.5 Long Term Influences of Funding Events

The presence of measurable effect by funding events on web-search traffic has been proven. It is of interest whether those events have long-term effects and how those effects might be characterise. Therefore, an analysis of long-term effects after funding events was conducted. This analysis focused on changes in the intercept, slope and variance after funding events. In the following, the results are discussed and related to the topic of this thesis.

6.5.1 Structural changes

The results in table 2 indicated that the fast majority of events do not cause any form of structural changes in the time series data. This observation stays persistent throughout different significance levels. Even at significance of $\alpha = 10\%$ more than 50% of all events display a coincident outcome of the dummy regression.

A further analysis considering the sequential position of the funding event did not reveal any logical pattern. Indeed, the results between significant and insignificant events display only minor differences. Considering that the group of significant events is slightly larger, the percentual distribution of regression outcomes among the two groups coincide. This raises an important question. Can an ignored event influence on the interest in the long term? The fact that some investment investment event did not cause a hype does not necessarily have the implication of no long-term influence in the interest.

As discussed in the chapter 2.1 VC investments are used by new ventures for all kinds of business activities. Therefore, any funding event must be seen either as the starting point of an marketing offensive, development of a new product or expansion into a new market. Thus, customer attention might be attracted over time. The instant hype caused by the venture capital community should not be seen as a prerequisite for a funding event to show effect in the long term.

Therefore, funding events cause in most of the cases a temporary outbreak in the interest for a new venture. As established in the previous sub-section, it is attention from the VC community attracted during funding events. The results in table 2 clearly prove that this short-term attention, the data underlying generating process does not change fundamentally. In terms of the VC community and customers of the new venture, one might conclude that as soon as the outbreak if over, the interest of customers be observed again. This is the group which drives the fundamental data generating process and which forms the fundamental value of every company.

However, there is a small number of events causing long-term effect on the interest of a new venture, which implies a long-term change in the customer value. Consider the outstanding amount of concurrent regressions even at high significance level. At a significance of $\alpha = 1\%$ and with a sample size of 294 one can expect 4 events be falsely classified as significant. The amount of 43 is outstanding. Therefore, the angle change of the new slope after the funding event was analysed. Concurrent and dissimilar regressions were used, as both imply a change in the slop. Figure 22 shows the distribution of the change in the angle of the

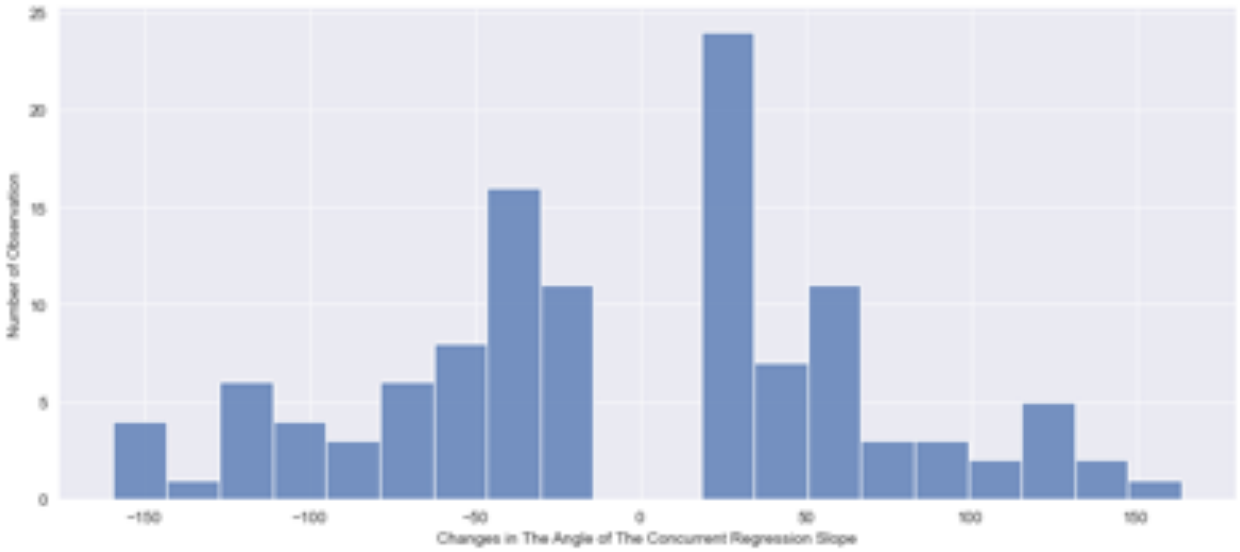

Figure 22: Angle Changes in Concurrent Regression

regressions. For every concurrent and dissimilar regression both slopes were taken and their slope angle was calculated. The slope angle of the first slope was compared to the slope angle of the second slope. The change in the angle was calculated for all such regression. One can observe that the distribution is close to normality, with the detail that values around 0 are missing. this is due to the fact that regressions with a slope angle change of 0 degrees are not concurrent nor dissimilar. Such regressions are either parallel or coincident. Due to the fact, that the slope angle changes are distributed similarly on both sides of the distribution, it needs to be concluded that funding events leading to concurrent or dissimilar structural changes have an ambiguous long-term influence on the slope. A funding event might cause a positive long-term effect on the interest but it is as likely to cause a negative long-term effect. A further investigation in has not revealed any logical pattern.

A similar analysis was conducted for parallel and dissimilar regressions to analyze, whether the change in the intercept is positive or negative. The results are similar as in figure 23. The resulting distribution is mirrored with no values around 0. Therefore, no logical pattern was found.

6.5.2 Changes in Variance

It was established that hype caused by investment activity is generally short. Long-term structural effects are observed rarely and their implications are ambiguous. It remains to discuss possible changes in the variance after funding events.

It is not trivial to relate changes in the variance of interest to the actual investment process. In the stock market fluctuations in price imply uncertainty. Variance is by definition a long-term metric. The variance of a process is defined over a certain sample or time interval. As discussed in the previous sub-chapter, outbreaks due to funding events display the interest of the VC community, while periods without funding events display the interest-generating process by customers and other individuals. I propose to think of variance increases as positive in the context of new ventures. Unstable variance implies a vivid process of interest generation. Unstable interest might be caused when a new venture expands to new markets and addresses a new group of customers. Since the new venture is not established in the new market, interest is displayed rather irregular and unstable. Following this logic, the interest in a new venture might be especially unstable in its early years, when no stable customer base. During the lifetime of a new venture, this customer base grows and new customers form only a small part of the overall interest.

In table 3 a decreasing trend of rejection over time of the null hypothesis for the Levene and Bartlett is observed. Not that table 3 displays funding events sequentially and by funding types. Both groups display the same declining rejection rate. This corresponds to the earlier assumption. Startups as early funding stages cause more often changes in the variance of interest. However, as soon as the startup grows and becomes a new venture, funding events stop causing instability in the interest-generating process. The F-test with $H_1 : \sigma_1^2 < \sigma_2^2$ clearly displays that the variance is growing over time, however stabilizes. The F-test with $H_1 : \sigma_1^2 > \sigma_2^2$ confirms the result. Early events do not lead to less, but more variance.

Considering figure 19 one can observe that the distributions of funding even with significant or insignificant outbreaks do not display major differences. Both types of events can cause long-term changes in the variance of interest. Therefore, funding events might also display long-term changes in terms of variance independent of whether they have been noticed by the VC community. As interest in funding events relates to customers and other individuals, which are attracted by the company's operation and products, it is concluded that all funding events influence the long-term interest generated by customers, independent of whether the VC community noted those funding events.

7 Conclusion

This thesis analyzed the effects of funding events on the web-search traffic related to technological based new ventures. The necessity of alternative data during the investment process has been theoretically derived. The predictive capabilities of web-search traffic have been presented. Web-search traffic was proposed to be further researched as alternative data, which might come in handy during the VC investment process. This proposal was founded on earlier findings by Malyy et al. (2021). The analysis in the thesis has undoubtedly found a strong influence of VC investment activity on web-search traffic related to new ventures. The strong statistical outbreaks have been shown to have a median p-value of 0.96. It has been shown that funding events display similar long-term effects on interest independent of whether their outbreaks are statistically significant. Funding series clearly follow a positive trend in outbreaks until series C rounds. Later down rounds receive less attention. Thus, web-search traffic clearly reacts to the venture capital investment process and mimics the theory of new venture valuation. Therefore, web-search traffic is strongly suggested to be further researched and implemented into the venture capital investment process.

I propose further research to focus on the link between valuation points and outbreaks in web-search traffic caused by funding events. More specifically, I propose to model the valuation of a new venture at the point when funding takes place. As endogenous variables might be considered: standard score at the funding event, type of funding event, venture capitalists involved in the investment process, and the position in the sequence of funding event.

Furthermore, a similar analysis like in this thesis might be repeated for another industry with large venture capital activity, like the biotech industry. It remains unknown, whether the results in this thesis are reproductive among different industries. Considering that technological-based new ventures provide products, which are available online this implies easy accessibility for possible customers. I assume that this accessibility is not present in industries like biotech. Therefore, web-search traffic might not display the same dynamics and insights as for technological-based new ventures.

As mentioned in chapter 3, internet activity is becoming a mirror of individual and group behavior. Due to the timely and easy accessibility of web-search traffic, it displays immediately new trends, desires, and needs. In an all-digital world, Google Trends might serve as a constant tracking device for any topic. I expect an immense increase in research and real-world applications of web-search traffic as alternative data. The potential of web-search traffic is by far not fully researched. This master thesis only scratched the surface.

Appendices

A Table of Companies Subject to Analysis

Company Name	Fud. Year	GT Code	Mean z-score	Med. p-value
Affectiva	2009	/m/0ql28cb	1.607585	0.860398
Affirm	2012	/g/11fkl4bl9w	1.055263	0.832268
Afiniti	2006	/g/11gv122lsp	1.092006	0.771532
AiCure	2010	/g/11b5yrrtk4	0.781636	0.470052
AiFi	2016	/g/11g7y_93_d	1.326235	0.835101
Algolia	2012	/m/010r912j	1.930884	0.956079
Algorithmia	2013	/g/11c52db9bk	2.226775	0.992087
ALICE Technologies	2013	/g/11f0207q3x	0.391057	0.725587
AMP Robotics	2015	/g/11c6t_4333	1.693938	0.973146
Anodot	2014	/g/11f01c8nl_	0.365014	0.519798
Appier	2012	/g/11f8p319wy	2.792354	0.998725
AppZen	2014	/g/11cpmhgv4_	1.471386	0.880670
Area 1 Security	2014	/g/11hd1s2xnf	2.396023	0.981637
Arterys	2011	/g/11fy2bbkyh	1.248272	0.928575
Atomwise	2012	/g/11dxpws3mz	1.325953	0.851670
Automation Anywhere	2016	/m/0b6fhdx	1.551365	0.962723
Babylon Health	2013	/g/11clgxrfm1	1.758142	0.968637
Behavox	2014	/g/11c57dpc5h	1.221355	0.758184
Benson Hill	2012	/g/11c73y8f0x	1.129416	0.881693
BioCatch	2010	/g/11dxq4ndlm	1.310773	0.925252
Bowery Farming	2014	/g/11dxp_hn78	1.626614	0.956052
BUILT Robotics	2016	/g/11fy21rpc9	1.323874	0.671288
Cape Analytics	2014	/g/11dx9gkqck	1.848766	0.994856
Casetext	2013	/g/11dxpjdhdb	1.233442	0.803861
Cerebras	2016	/g/11g9mm3yd5	2.037160	0.950617
Citrine Informatics	2013	/g/11c73stlbz	0.935329	0.822531
CognitiveScale	2013	/g/11cm0d6sz4	1.252747	0.788246
ComplyAdvantage	2014	/g/11dxpkf339	1.350133	0.948465
Conversica	2007	/g/11h59s3sw3	0.995929	0.786709
Covariant	2017	/g/11ngkwl2fd	1.128643	0.796583
Coveo	2012	/m/03qkfwk	2.919886	0.998833

CrowdStrike	2011	/g/11bz0yw54s	2.631619	0.995692
Cybereason	2012	/g/11dyzf9js9	2.556323	0.990762
Darktrace	2013	/g/11c2l23pjr	2.384152	0.996479
Dataiku	2013	/g/11bzyqdcsn	2.497967	0.997369
DataRobot	2012	/g/11f017ds55	1.642249	0.938497
DataVisor	2013	/g/11bxg16v_z	1.554825	0.918997
DeepMap	2016	/g/11j43f60ms	2.380604	0.973739
Dynamic Yield	2011	/g/11h7c757qy	1.834856	0.985607
ExaWizards	2016	/g/11f8jt1zbm	0.619348	0.662526
Featurespace	2011	/g/11c73nt4x0	1.971577	0.944789
Fiddler Labs	2018	/g/11fjzxn2nc	1.002543	0.816142
Fortem Technologies	2016	/g/11c52tj97g	2.158735	0.988291
Freenome	2014	/g/11cp7rlv9h	2.005543	0.972614
Grabango	2016	/g/11h91z4vlg	2.040891	0.996017
Graphcore	2016	/g/11f006dwxx	2.476056	0.998980
Habana Labs	2016	/g/11fqsx2nv1	2.525180	0.977381
Hacarus	2014	/g/11c73vp35y	0.759028	0.762981
Healthyio	2013	/g/11fhqkq_b6	1.069949	0.959157
Horizon Robotics	2015	/g/11fy1rflmk	1.217538	0.904610
Hyperscience	2014	/g/11clsrqbq8	0.849411	0.842475
Inceptio Technology	2018	/g/11fkczmw97	0.953084	0.857919
Insilico Medicine	2014	/g/11fy_q8yf8	1.728810	0.944751
insitro	2017	/g/11gjx89qyc	3.042139	0.999483
Jina AI	2020	/g/11k6f81lmq	1.127124	0.950720
Kindred Systems	2014	/m/09mw4yy	3.092142	0.999289
Kneron	2015	/g/11c5r0m21k	1.581330	0.966316
KONUX	2014	/g/11gxt5j1wg	1.486619	0.955951
Landing AI	2017	/g/11hb2stpqb	1.647303	0.876084
LawGeex	2014	/g/11c742kymd	1.122737	0.907047
LeanTaaS	2010	/g/11b_24kgtk	1.562317	0.913611
Lemonade	2015	/g/11g9q62bm6	1.678217	0.977235
MEGVII	2011	/g/11fn23jh37	1.112533	0.871990
Nauto	2015	/g/11dy1mfvb4	1.264317	0.887192
Neurala	2012	/g/11bwn772bh	1.160197	0.892630
NotCo	2013	/g/11fy28js0s	1.714268	0.882805
Nuro	2016	/g/11g18_fthf	1.923694	0.967805
Obsidian Security	2017	/g/12611sqrk	2.594190	0.981832
ONEFLOW	2017	/g/11cn5kwmmj	1.125591	0.854713

Onfido	2012	/g/11c1wr530_	1.808587	0.978474
Orbital Insight	2013	/g/11hdvdrxsd	1.740174	0.965831
Osaro	2015	/g/11bwkg0tqx	1.644727	0.949965
Outrider	2017	/g/11gxn1t1cc	1.781191	0.950354
Owkin	2016	/g/11dxq34t0c	1.861033	0.977736
PAIGEAI	2017	/g/11f6cqgnz8	1.955829	0.981177
PerimeterX	2014	/g/11g9mlhkj6	2.238324	0.985219
Petuum	2016	/g/11f62wywqf	2.810182	0.992773
PolyAI	2017	/g/11fj8fxzgr	1.600990	0.948245
Ponyai	2016	/g/11j1_jkbw8	2.177530	0.989106
Preferred Networks	2014	/g/11hcw7x0wn	1.489236	0.950427
Primer	2015	/m/05x0v	0.186339	0.704190
Qventus	2012	/g/11dxpmgvzg	1.089477	0.874809
Rossum	2017	/g/11h1y90t3y	2.745732	0.994530
SenseTime	2014	/g/11f55596ch	1.721915	0.937342
SentinelOne	2013	/m/012r5lnd	1.272348	0.922203
Shift Technology	2014	/g/11g9mt7j38	1.226915	0.888345
Sift	2011	/g/11b7tc0m2r	1.090318	0.708779
Signifyd	2011	/g/11bxfqwwms	1.779351	0.983259
SigOpt	2014	/g/11f011lqt_	1.936278	0.927467
Snyk	2015	/g/11fy24_1tx	1.612632	0.921929
Socure	2012	/g/11c1vj86m4	1.594232	0.943851
SparkCognition	2013	/g/11cn9qfb6b	1.846310	0.991036
Sportlogiq	2015	/g/11c67x416d	0.727527	0.592937
Standard Cognition	2017	/g/11f259lw0r	2.152456	0.988868
StormForge	2015	/g/11h06m_1qf	0.628220	0.645643
Subtle Medical	2017	/g/11f7pbn83t	1.104006	0.917128
SuperAnnotate	2018	/g/11pwd0g5ph	1.460362	0.919211
Syte	2015	/g/11f765hc0d	1.426499	0.946476
Tachyus	2014	/g/11dxpmnlsj	0.631136	0.500000
Tactai	2012	/g/11dfk4rkcy	1.388341	0.963589
Text IQ	2014	/g/11f_j3vb1r	1.899534	0.979192
Textio	2014	/g/11cjk6qyvg	1.475583	0.931556
Theator	2015	/g/11h40yrx96	0.547185	0.576366
Tomorrowio	2015	/g/11dftbqwfz	0.979717	0.831817
Tractable	2014	/g/11dxq0t91c	2.018245	0.993588
TuSimple	2016	/g/11r6y6z60m	1.419878	0.902459
UBTECH Robotics	2012	/g/11hydk79t2	0.181853	0.487028

UiPath	2005	/g/11f50xjkr9	1.443179	0.915679
Unbabel	2013	/m/0138nb9l	1.760259	0.967446
Versive	2012	/m/0x0v3c7	2.009835	0.990019
Weights & Biases	2017	/g/11f7h7qw7s	1.847615	0.979205
WorkFusion	2011	/g/11b8v9r30y	0.968857	0.853728
Zestyai	2015	/g/11f7r14rvw	0.577333	0.362794
Zymergen	2013	/g/11dymfxl0_	2.246324	0.995443

References

Boonpeng, S. and Jeatrakul, P. (2016). Decision support system for investing in stock market by using OAA-neural network. In *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*. IEEE.

CHOI, H. and VARIAN, H. (2012). Predicting the present with google trends. *Economic Record*, 88:2–9.

Cochrane, J. H. (2005). The risk and return of venture capital. *Journal of Financial Economics*, 75(1):3–52.

DataReportal (2021). Digital 2021: July global statshot report. https://datareportal.com/reports/digital-2021-october-global-statshot.

Djeundje, V. B., Crook, J., Calabrese, R., and Hamid, M. (2021). Enhancing credit scoring with alternative data. *Expert Systems with Applications*, 163:113766.

DUN & BRADSTREET (2017). Alternative data: The hidden source of alpha. https://www.dnb.com/content/dam/english/dnb-solutions/alternative-data-the-hidden-source-of-alpha-whitepaper.pdf. Accessed: 2022-01-31.

Entrepreneur (2016). 5 ways to get media coverage as a startup. https://www.entrepreneur.com/article/274887. Accessed: 2022-03-10.

Fessl, A., Apaolaza, A., Gledson, A., Pammer-Schindler, V., and Vigo, M. (2019). "mirror, mirror on my search...": Data-driven reflection and experimentation with search behaviour. In *Lecture Notes in Computer Science*, pages 83–97. Springer International Publishing.

Fiancial Times (2020). Hedge funds scour alternative data for edge on covid and economy. https://www.ft.com/content/8d194207-f6bf-4dde-b0fe-93cb85dfb8a0. Accessed: 2022-01-31.

Forbes (2019). Alternative data. what is it, who uses it and why is it interesting? https://www.forbes.com/sites/forbesinsights/2019/12/12/alternative-data-what-is-it-who-uses-it-and-why-is-it-interesting/?sh=14425dde6123). Accessed: 2022-01-31.

Greenwich Associates (2018). Seismic shifts - the future of investment research. http://smallake.kr/wp-content/uploads/2018/12/thomson-reuters-and-greenwich-associates.pdf. Accessed: 2022-01-21.

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

Jagtiani, J. and 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.

Johnson, J. (2021). Coronavirus: impact on online usage in the u.s. - statistics & facts. https://www.statista.com/topics/6241/coronavirus-impact-on-online-usage-in-the-us/topicHeader$_{wrapper.\,Accessed:2022-03-06.}$

J.P. Morgan Global Quantitative & Derivatives Strategy Team (2017). Big data and ai strategies - machine learning and alternative data approach to investing. https://web.archive.org/web/20180722070124/https://www.ravenpack.com/research/jp-morgan-big-data-ai-machine-learning-alternative-data/. Accessed: 2022-01-16.

Kaplan, S. and Lerner, J. (2016). Venture capital data: Opportunities and challenges. Technical report.

Katherine Noyes (2016). 5 things you need to know about data exhaust. https://www.computerworld.com/article/3070475/5-things-you-need-to-know-about-data-exhaust.html. Accessed: 2022-01-31.

Kjartan Rist (2020). Vc's outsized economic impact will power a new golden age. https://www.forbes.com/sites/kjartanrist/2021/07/28/vcs-outsized-economic-impact-will-power-a-new-golden-age/?sh=4110704a3204. Accessed: 2022-03-10.

Kraut, R. and Burke, M. (2015). Internet use and psychological well-being. *Communications of the ACM*, 58(12):94–100.

Lechler, T. (2001). Social interaction: A determinant of entrepreneurial team venture success. *Small Business Economics*, 16(4):263–278.

Li, X., Law, R., Xie, G., and Wang, S. (2021). Review of tourism forecasting research with internet data. *Tourism Management*, 83:104245.

Maggio, M. D., Ratnadiwakara, D., and Carmichael, D. (2022). Invisible primes: Fintech lending with alternative data. Technical report.

Malyy, M., Tekic, Z., and Podladchikova, T. (2021). The value of big data for analyzing growth dynamics of technology-based new ventures. *Technological Forecasting and Social Change*, 169:120794.

Mario Gabriele (2021). Tiger global: How to win. https://www.readthegeneralist.com/briefing/tiger-global. Accessed: 2022-01-21.

Matt Monday (2018). The hidden steps to startup success. https://www.forbes.com/sites/forbestechcouncil/2018/01/05/the-hidden-steps-to-startup-success/?sh=68bfa9bc64f4. Accessed: 2022-03-10.

Michael Peregrine (2021). An overview of the ipo process. https://www.forbes.com/sites/michaelperegrine/2021/09/15/preparing-to-go-public-an-overview-of-the-ipo-process/?sh=24a451a16931. Accessed: 2022-03-10.

Miloud, T., Aspelund, A., and Cabrol, M. (2012). Startup valuation by venture capitalists: an empirical study. *Venture Capital*, 14(2-3):151–174.

Monika and Sharma, A. (2015). Venture capitalists' investment decision criteria for new ventures: A review. *Procedia - Social and Behavioral Sciences*, 189:465–470.

National Venture Capital Association (2019). Venture monitor q2 2019. https://files.pitchbook.com/website/files/pdf/2Q_2019_PitchBook_NVCA_Venture_Monitor.pdf. Accessed: 2022-01-06.

Oracle (2019). Searching for alpha-using alternative data sets & new gen technologies. https://www.oracle.com/a/ocom/docs/industries/financial-services/searching-alpha-alternate-data-sets-wp.pdf. Accessed: 2022-01-31.

Óskarsdóttir, M., Bravo, C., Sarraute, C., Vanthienen, J., and Baesens, B. (2019). The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. *Applied Soft Computing*, 74:26–39.

Pai, P.-F., Hong, L.-C., and Lin, K.-P. (2018). Using internet search trends and historical trading data for predicting stock markets by the least squares support vector regression model. *Computational Intelligence and Neuroscience*, 2018:1–15.

Painter, M. (2018). Unlevelling the playing field: The investment value and capital market consequences of alternative data. *SSRN Electronic Journal*.

Raven Pack (2016). Data hoarding and alternative data in finance - how to overcome the challenges. https://www.ravenpack.com/blog/data-hoarding-alternative-data-finance/. Accessed: 2022-01-31.

Sandberg, W. R. and Hofer, C. W. (1987). Improving new venture performance: The role of strategy, industry structure, and the entrepreneur. *Journal of Business Venturing*, 2(1):5–28.

Sergio Paluch (2021). Top angel investors – the definitive list for 2021. https://betaboom.com/blog/top-angel-investors/. Accessed: 2022-01-15.

Stampfl, G., Prügl, R., and Osterloh, V. (2013). An explorative model of business model scalability. *International Journal of Product Development*, 18(3/4):226.

Techopedia (2019). What does data exhaust mean? https://www.techopedia.com/definition/30319/data-exhaust. Accessed: 2022-01-31.

Terry Flanagan (2016). 'early days' for alternative data. https://www.marketsmedia.com/early-days-alternative-data/. Accessed: 2022-01-31.

The Business Research Company (2021). Alternative data global market report 2021: Covid-19 implications and growth. Technical Report SKU-BRC16754172, The Business Research Company, The Business Research Company, US.

Vara, W. P. (2013). Risk-based new venture valuation technique: Win-win for entrepreneur and investor. *Journal of Business Valuation and Economic Loss Analysis*, 8(1):1–26.

Varian, H. R. and Choi, H. (2009). Predicting the present with google trends. *SSRN Electronic Journal*.

Vishal Persaud (2021). Thow tiger global threw out the venture-capital rulebook and bulldozed the competition in 2021. https://www.businessinsider.com/tiger-global-changed-venture-capital-startups-funding. Accessed: 2022-01-15.

Vladimirovich, V. K., Grigorievna, A. P., and Shalaev, V. (2015). An analysis of the impact of venture capital investment on economic growth and innovation: Evidence from the USA and russia. *Ekonomski anali*, 60(207):7–37.

Wikimedia Foundation (2022). Alternative data (finance). https://en.wikipedia.org/wiki/Alternative$_d ata_($finance). *Accessed* : 2022 − 01 − 31.

List of Figures

List of Tables