MSc Data Science for Business

Research Paper

Academic Year 2021 - 2022

Using data science to enhance Venture Capital deals

Overview of the data science field in the Venture Capital industry

Louis Bertolotti

Under the supervision of

Prof. Jean-Edouard Colliard

Contents

1 Preface and acknowledgment

Back in 2017 when I started my first year at HEC Paris just after my "classe prépa", I failed to connect with my finance courses. I was doing the minimum to validate my courses and didn't think much of it. After a gap year in data science, I once again gained an interest in finance as I saw the potential of existing datasets, were the data to be centralized in data warehouses instead of being spread across multiples spreadsheets.

More and more intrigued by the relatively obscure links between finance and the advanced data processing techniques that I developed during my two years in the Data Science for Business master, I decided to follow Jean-Edouard Colliard's courses of finance. They were very different from my L3 courses, and managed to clearly get my attention for their mix of strong theoretical bases and very applied exercises.

One of those courses evoked the growing use of data science in the Venture Capital world. This caught my attention as I have a very strong interest in the tech industry, and had only one regret during my internships in HEC, which is to not have had the opportunity to have an internship in a VC fund. I decided to choose this subject since it enabled me to both discover the whole VC ecosystem and to explore a quite uncharted aspect of finance and data science. Furthermore, Maxime Bonelli, a PhD student at HEC Paris, gave a presentation during one of my finance courses on this very subject. After exchanging with him, we agreed to cooperate on a survey of the maturity of data science in the Venture Capital industry.

I would like to thanks Jean-Edouard Colliard for the quality of his finance courses and for his help in sizing and orientating my research paper. I would also want to thanks Maxime Bonelli for his crucial help in the design of the survey and in the data gathering which made it possible. I would like to thanks all of the data scientists and VC investors who took the time to answer this survey on their free time. Finally, thanks to all of the DSB alumni who helped me on the survey or guided me to the right people when I asked them.

2 Executive summary

In this research paper, I try to show why data science techniques are more and more adopted by Venture Capital firms. Improving sourcing, screening and reducing bias through behavioral science are key ideas to bolster the investment process. In other words, data science helps to extend the pool of companies to select good investments from in less time than before.

By directly surveying data science actors working in Venture Capital firms, I built an overview of the maturity of the field, from the technologies used to the perceived effect on investments. Crossing these answers with financial data shows that there are different types of Venture Capital firms using data, from early experimentation firms to full-fledged data product companies.

Notably, big firms may experience resistance to change, especially when they tackle multiple stages of the investment process. Late stage VC firms should instead be more careful of the success of data projects carried out with their portfolio companies.

Finally, it is possible to start a data science from scratch and to iterate until the firm is able to build intuitive internal products. This however requires good communication between the investment team and the data science team, which may require a profile with both a business and a technical formation.

3 Introduction

The Venture Capital investment industry has gone through countless transformations, especially in the last 20 years where the number of software start-ups have skyrocketed. [For] At the same time, data science was emerging as a discipline at the intersection of software, mathematics and business knowledge. Big datasets have enabled the use of machine learning techniques which are able to predict financial insights accurately.

Yet while some parts of finance such as hedge funds have slowly embraced algorithms, Venture Capital doesn't seem as advanced in terms of data science use. Indeed, this world is known for strong investor heuristics, who are used to meeting founders and to interact with humans to make the final investment decision. It seems at first glance that algorithms modelisation could prove difficult with so much human interaction in the process.

However, more and more VC firms announce that they are using data science products at various stages of the investment process. In order to map this emerging field, performing a survey of key individuals working in the field is the only way to understand why, how and where they used data science, since this information is not really available on public sources.

Based on various readings, my hypotheses are that sourcing and screening are key parts of the current data science techniques used in the industry.

What is the current maturity of the data science field in the VC industry ? How are they helping the deal process ? Is it possible to identify profiles of actors in the sector ?

4 Theoretical approach : using data science to gain an edge on the market

4.1 Understanding the VC investment approach

4.1.1 Start-up investment stages

A start-up needs cash to survive in the long term. While some of them can stay self-sustaining with the right business model, most of them requires investments to pay for top-tier talents, offices and various costs. The rise in the number of tech start-ups has increased this phenomenon : the SaaS (Software as a Service) model indeed promises future regular revenues once the product becomes mature enough and when it succeeds in attracting enough customers. A successful start-up needs to combine a vision of the product, and a good product-market fit in order to succeed.

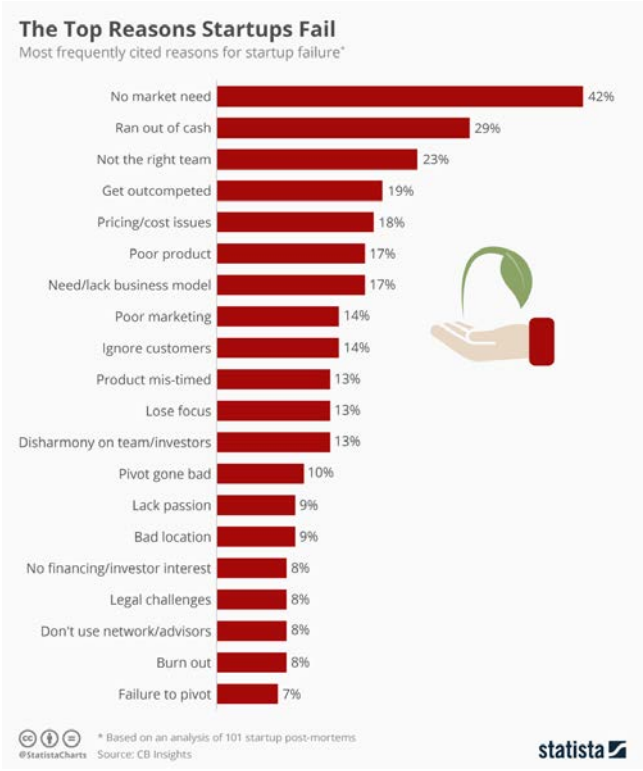


Figure 1: Main reasons behind the failure of start-ups (Source : Statista)

In order to better understand the way VC investors approach investment in a start-up, a distinction must be drawn between Venture Capital and Private Equity funds. Private Equity investment aims to invest in already mature companies, which most of the time have management or organisational issues. By taking a majority stake in the company, the Venture Capital firm tries to impulse necessary changes in order to increase the company's profit. LBO (Leveraged Buy-out) is an example of such a strategy, where debt is raised in order to acquire a majority stake and to set drastic objectives for the management.

Venture Capital funds are special since they carry more risk [Neu19] than private equity investments, hoping in turn for a greater return on investment. The pools of investment candidates is also more difficult to apprehend. There are multiple stages where investors can enter the capital of a company [RME22] :

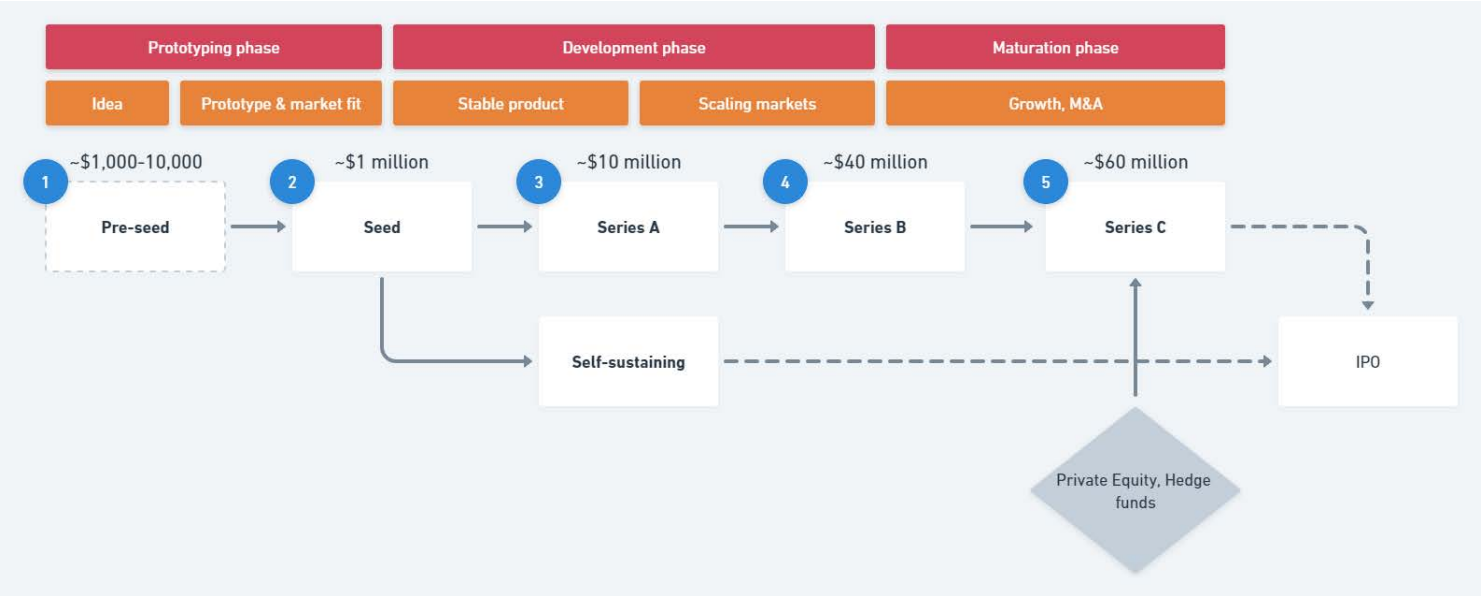


Figure 2: Start-up investment flowchart (Source : Investopedia article, original chart)

- **Pre-seed** : this phase is optional as shown on the chart. It corresponds to the part where the start-up is still just an idea in the head of its co-founders, or when it barely started prototyping. It is not really considered as the first investment in the "legal" structure of the company. This would cover the costs of the initial software or machines necessary to build a prototype before the start-up is launched.

- **Seed** : this is the first "official" wave of financing that a company can pretend to. Usually, the idea is clearly defined at this point, and the start-up acquires a legal status with equity injected by the co-founders. Two other actors can then intervene : an accelerator such as Y Combinator which will help the co-founders to ensure that the idea and the product-market fit are strong, and "angel investors" who inject equity very early when the product is barely tangible. The accelerator also usually invests money in the most promising projects. The median of seed projects amounted to $1 million in 2021. [RME22]

- **Series A** : the Series A is performed on projects which already have a working prototype and some clients. They usually come as a confirmation that the product-market fit was accomplished through key metrics (retention, MRR for SaaS...). While the company is often far from being profitable at this point, it has shown that it can expand. The money is therefore used to grow the size of the team or to implement teams in foreign markets to try to export the successful formula. The median of Series A projects amounted to $10 million in 2021. [RME22]

- **Series B** : the series B comes as an improvement of the Series A funding, for start-ups which have shown their ability to replicate their success on other markets. At this point, the equity is more and more diluted, and the VC investors invest with less risk (and therefore with the promise of fewer returns). The objective of this funding is to keep the expansion going and to start thinking about moves on competitors with potential acquisitions. The median of Series B projects amounted to $40 million in 2021. [RME22]

- **Series C** : Series C are the final stage of the VC investment process. At this point, the start-up has turned into a scale-up which is no longer simply a Venture Capital investment target. Private Equity

funds and Hedge funds start to also invest in the asset, as management inefficiencies may have settled in, or synergies may be available on the market. The median of Series C projects amounted to $60 million in 2021.[RME22]

- **IPO** : this phase is optional. If the company is successful enough, it may decide to enter an Initial Public Offering (IPO), which turns the company public by selling share on the stock market. This may also happen directly if the business model of a start-up makes it possible to immediately be profitable. By investing correctly, those companies can grow with little dilution and end up going for an IPO.

Another trend in Venture Capital is the development of Revenue-Based Financing. The idea is that the investors lend money to the start-up : it has to refund 3 to 5 times this amount, but not on a regular basis. The investor takes a percentage of the revenues until the amount is refunded. This prevents dilution while still making money available for the development of the start-up, without the debt risks for the co-founders.

4.1.2 VC firm investment process

Now that we have seen when a VC firm can invest, let us see how it invests. The investment logic in Venture Capital is comparable to a funnel since there is a large pool of start-ups to select from, which will be thinner and thinner after each step.

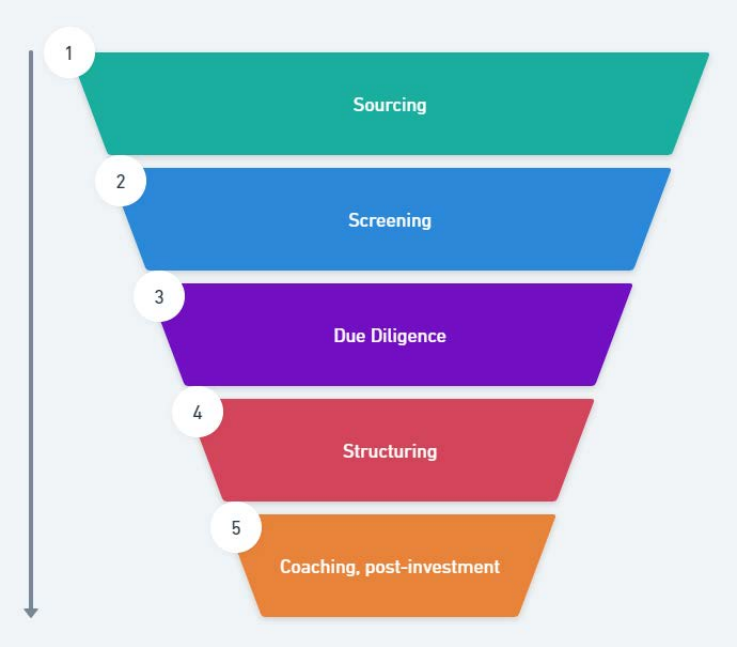


Figure 3: VC investment funnel (Source : Maxime Bonelli, original chart)

- **Sourcing** : during this phase, the fund tries to amass as many start-up contacts as possible. In fact, it will even try to interview competitors working a relatively similar product in a very similar area in order to find the most promising performer. VCs will go to dedicated conferences, organize pitching events and use their network in order to find those targets. The idea is to reach a certain volume without filtering too much at this point. [Kap17]

- **Screening** : this phase of the pipeline aims to apply a first layer of filtering on the start-ups. The process can be loosely organized with simple conversations on a company or they can follow a thorough evaluation process involving partners pitching and writing notes to their

leadership. Here, investors will often try to favour fields and regions were they have respectively experience and a strong network of contacts.

Another key point is the investment thesis of the fund, which sums up the direction that is it taking when investing. Is it going to invest in a specific region or on an entire continent ? Is is software-orientated or specialized in healthcare ? Does it have a goal to bolster the growth of CSR-compliant companies to fight against climate change ? The DNA of the fund is necessarily going to affect the screening process in order to tailor the final list of start-ups to the fund's vision.

Finally, certain funds only target certain stages. Seed funds will try to invest in a lot of companies while carrying a high risk per investment, while others will only invest in series A, B or even C stages. However, it is common for early investors to enter further rounds as a way to keep securing a successful investment.

- **Due Diligence** : during the due diligence, the fund will gather more precise information in order to evaluate the health and potential of the companies. The due diligence can be performed by analysts from the fund and by analysts from specialized strategy consulting firms. The fund will ask detailed questions during weekly meetings with the consulting firms, which they will answer to with slide describing the state of the market, the maturity of the companies and a certain amount of financial KPIs. The start-up will usually also produce its own due diligence, called a seller side due diligence, while the fund's one is called a buyer due diligence.

- **Structuring** : the structuring is a phase where the fund and the start-up will negotiate the conditions of the deal. The crux of the matter is the valuation of the company, which will influence the amount of money a founder can expect when selling his shares. The negotiation will also determine how much equity is gained by the fund in exchange for a fixed investment. This phase usually includes a "liquidation preference" : as long as the shareholders have not received the amount of their investment back, they have a priority when it comes to buying new shares over classical investors. It also includes superior voting rights

which enables to fund to pilot the strategy of the start-up alongside co-founders. Finally, a "ratchet" clause can be included, which protects the equity owned by the fund from later dilutions.

- **Post-investment** : once the deal is concluded, the fund will try to leverage its investment by providing as much coaching as possible to the founders. Usually, partners will use their field knowledge to advise the board of the company on strategic decisions. Most funds also have a dedicated group of experts who usually have first-hand experience on the creation of start-ups : they can be software engineers, CEOs, or academics who will bring state-of-the-art advices to the board. And some funds even deploy entire operations teams who specialize in disciplines like marketing or data science in order to help the start-up on a daily basis. This method is very efficient as it creates an unfair advantage over competitors who do not have access to this type of advice, while reducing the risk for the fund by ensuring that the start-up generate continous growth on the long term.

4.2 Increasing performance thanks to data science techniques

This state of the market where investments are increasing is coupled with increased competition between VC firms. As both the total amount of money invested and the number of deals drastically increases, more investors want to join the fray.

Therefore, VC firms need to develop a differentiating factor in order to find the best deals. While the war for top talents is one aspect of this strategy, it is not enough to tackle the sheer amount of start-ups to process at any stage of the investment process. Another issue is the lack of investment data during the early stages of a start-up: the start-up is not making profits yet, so financial performance doesn't make much sense here. [CBC21] Instead, data scientists will try to find alternate data able to give insights on the quality of the team.

There are multiple stages where a VC firm can use data science to increase its productivity :

- Automatize sourcing : the VC firm uses multiple scrapping sources such as LinkedIn or GitHub accounts to find new start-up targets;

- Screen companies faster: the VC firm evaluates multiple companies automatically as a way to ease the due diligence work;

- Coach the companies and give them "unfair" advantages: by tying a close relationship to the company, the fund can use its expertise to guide the management and it can provide experts in their domain, such as biotechnology, data science or marketing.

For instance, here is how Google Ventures used algorithms back in 2011 to screen profiles :

> "To make its picks, the company has built computer algorithms using data from past venture investments and academic literature. For example, for individual companies, Google enters data about how long the founders worked on start-ups before raising money and whether the founders successfully started companies in the past. It runs similar information about potential investments through the algorithms to get a red, yellow or green light. [...] "A lot of times V.C.'s will say, 'We're not just money, we're value-add,' and I've always been somewhat doubtful of those claims," Mr. Walker said. "With Google Ventures, those claims are completely justified." " [Mil11]

Interestingly, this article also shows that breaking bias in investment is a good way to find start-ups that would not have been noticed otherwise : start-ups which are headquartered in another area than the VC firm headquarter are more successful on average because they are submitted to less heuristics. The firm can also closely help the start-up by providing key infrastructure and advice from experts. This is a point raised in [Blo+22] : VC investors affected by strong bias are beaten by ML algorithms, while only seasoned experts who reduced their investing bias managed to beat the algorithms.

4.3 Behavioral science

To further elaborate on this point, it is necessary to dwell into behavioral science. In the case of VC funds, an interesting insight is the fact that the people which are part of an early start-up are a key economic component of the future success of the company.

Some of the funds existing on the market employ behavior scientists instead of data scientists. They often have a formation in social science or neuroscience along with computer science or statistics skills. Start-ups are also emerging in this field in order to provide this kind of analyses to VC funds.

I had the opportunity to interview Mathieu Nasri, the co-founder of the start-up Unfair. This company aims to enhance VC investment by both reducing biases on the investment side and by ensuring that the co-founders of a start-up are a good match in terms of personality. It was founded 6 years ago on research about which parts of the brain determine success. Unfair derived a method called the OBA (Online Brain Analysis) : the person being evaluated has to complete a form with questions about concrete situations, and the results are summed up in an automated psychological evaluation. It automatically provides a report showing the main advantages and roadblocks that those personalities will inject in the start-up project.

Unfair started on a sample of 200 founders to train a model able to evaluate their personality, extracted insights about the link between these personality traits and success and compared the results of their model to the success of past start-ups. Their model was able to predict with a 89% accuracy whether the start-up would fail. It was developed jointly with 8 VC funds in order to gather the necessary data. [NG]

This system is all the more useful as it handles the relationship between members of a team. It can predict whether two co-founders complete each other correctly. It benchmarks their traits and depending on the profiles, it can detect unique flaws in the relationship emerging from this single combination of traits. Meetic's founder, Marc Simoncini, explained that Philippe Chainieux was vital in order to manage the day-to-day operations of his start-ups : he recognized that he was more of an idea creator than a manager. Chainieux fit this role as his personality was very different from Marc's, who considers

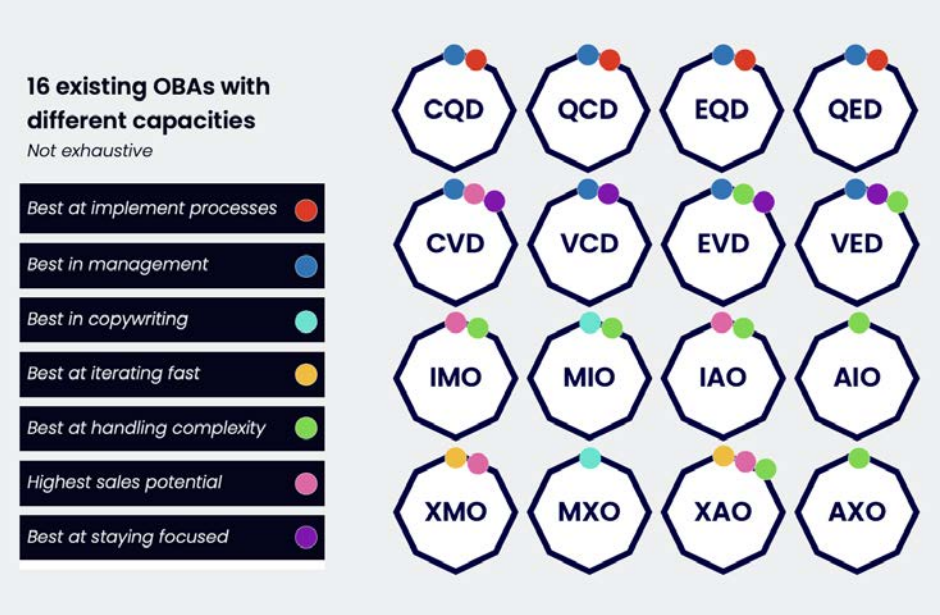


Figure 4: Various profiles produced by the OBAs (Source : Unfair website)

that he is good at attracting good profiles to his companies.

On the other hand, the platform is also able to determine behavioral biases in investors' decisions. Investors profiles are subject to investment biases, as they will answer differently to co-founders personalities. While some favour honest and humble profiles, others will prefer more adventurous personalities. An investment teams with unbalanced personalities will be biased in most of its investment decisions, adding risk to the investment profiles.

Furthermore, Unfair offers consulting services in order to react to those reports. While changing one's personality is very difficult, it is possible to "balance" a team with new profiles which complement the others. For instance, the founders can recruit a complementary profile as a first employee, while VC funds can recruit investors with different personality traits. A key point is that those traits can be independent from the job or specialization of the new employee : some engineers recruited by VC funds are not as receptive to rational presentations and react more to emotions, which can be counter intuitive. Unfair's product can then be seen as a second selection

system once multiple profiles are detected for a job.

5 Data science techniques used on the field

5.1 Designing the survey

5.1.1 Motivations

Now that we have evaluated the reasons why data science can enhance various investment phases of the VC funnel, we can take a look at the way those techniques are actually used on the field. The best way to proceed was to directly survey the funds which were the most likely to use data science. I worked with Maxime Bonelli, a PhD student in Finance at HEC Paris, in order to design it and ask the right questions to the funds.

By directly polling members of data science teams working in Venture Capital funds, we could gather information on the maturity of the data stack while also receiving precious insights on the perceived efficiency of those techniques from within the fund. This information is really hard to come by in traditional data sources such as Statista or Refinitiv where there is no direct data on the the use of data science by a fund.

Another reason why I accepted to perform the survey was that it could benefit Maxime's working paper on the performance of Venture Capital Funds. The survey would kill two birds in one stone and help to see whether the results of his paper applied to existing VC firms.

5.1.2 Methodology

Maxime gave me access to a 176-long list of VC funds which appeared to be using data science according to the LinkedIn profiles of their employees. Most of them where located in Europe, in the US or in London.

Our objective with this survey was two-fold :

- **Quantitative** : knowing how many funds were actively using data science techniques;
- **Qualitative** : understanding how and when those techniques were used in the deal flow.

Therefore, our questions try to evaluate those two aspects in terms of intensity and precision. The last question is an invitation for an in-depth interview in order to perform a detailled analysis of the techniques used.

Key aspects which were polled are :

- Whether a fund is using data science, with a cursor on the level of use;
- The types of techniques used in the fund;
- How those techniques are applied in the fund (when, where, how);
- Whether those techniques are adopted smoothly or if resistances exist.

Here are the exact questions which were asked :

- 1 - To which extent are you using data science techniques in your fund ? (*multi-choice question*)
- 2 - Are you planning to deploy such technologies or to expand their use in the next 12 months ?
- 3 - If you use such technologies, during which steps of the investing process do you use them ? (*multichoice question*)
- 4 - What are the main reasons driving your use of these technologies ? (*multi-choice question*)
- 5 - Which types of data sources are you using ? (*multi-choice question*)
- 6 - Which technologies are you using (algorithms types, programming languages/packages, on-board solution) ? (*multi-choice question*)
- 7 - Do you consider that those technologies helped you to enhance the quality of your start-up investments ?
- 8 - If they did, how ? (*multi-choice question*)

- 9 - How long did it take to implement those solutions internally ?
- 10 - Do you encounter resistance to the implementation of data science techniques or to automatisation ?
- 11 - Do you manage to find/recruit profiles who both master the business side and the technical side of data science implementation ?
- 12 - Would you be willing to be contacted again for a brief interview in order to better understand the technologies that you are using ?

5.1.3 Hypotheses

Before starting the survey, I drew some hypotheses on what the results would be in order to direct my questions during the interviews and to prioritize some analyses over others.

- **Recruitment** : Considering the difficulty to find profiles mastering both the data science technical side and the more business-orientated side inherent to VC funds, one of my assumptions was that most funds would find it difficult to find such profiles.
- **Technical stack** : I expected to find two main profiles of funds, those using MLOps products such as Dataiku or DataRobot which enable a few individuals to cover most of the data treatment steps quickly, and organized data teams developing internal tools for the fund.
- **Internal resistance** : I learned from my finance and data courses and from the experience of friends working in Mergers & Acquisitions that the finance world experiences some friction when it comes to adopting new technologies or automation systems. Even though Venture Capital inventors experience by design a closer link to companies developing these types of products, I expected internal resistance to adoption by managers.
- **Performance** : judging from successful use cases presented by some funds (especially in California), I expected increased financial revenues by funds using data science compared to normal funds. In particular, I

expected a higher number of deals performed by number of employees due to the productivity gains obtained during some steps such as the sourcing or screening.

5.2 Carrying out the survey

5.2.1 Response rate

Out of the 177 initial funds, I identified 106 funds which were employing data scientists or at least partners with experience in data science. The rest had probably externalised their work to build their platform or their employees hid their LinkedIn profile. I had the time to contact 97 of those funds, since the last ones were usually big insurance groups or banks with more than 2,000 employees : it was therefore quite difficult to determine if the data scientists were working with the venture capital arm of the group.

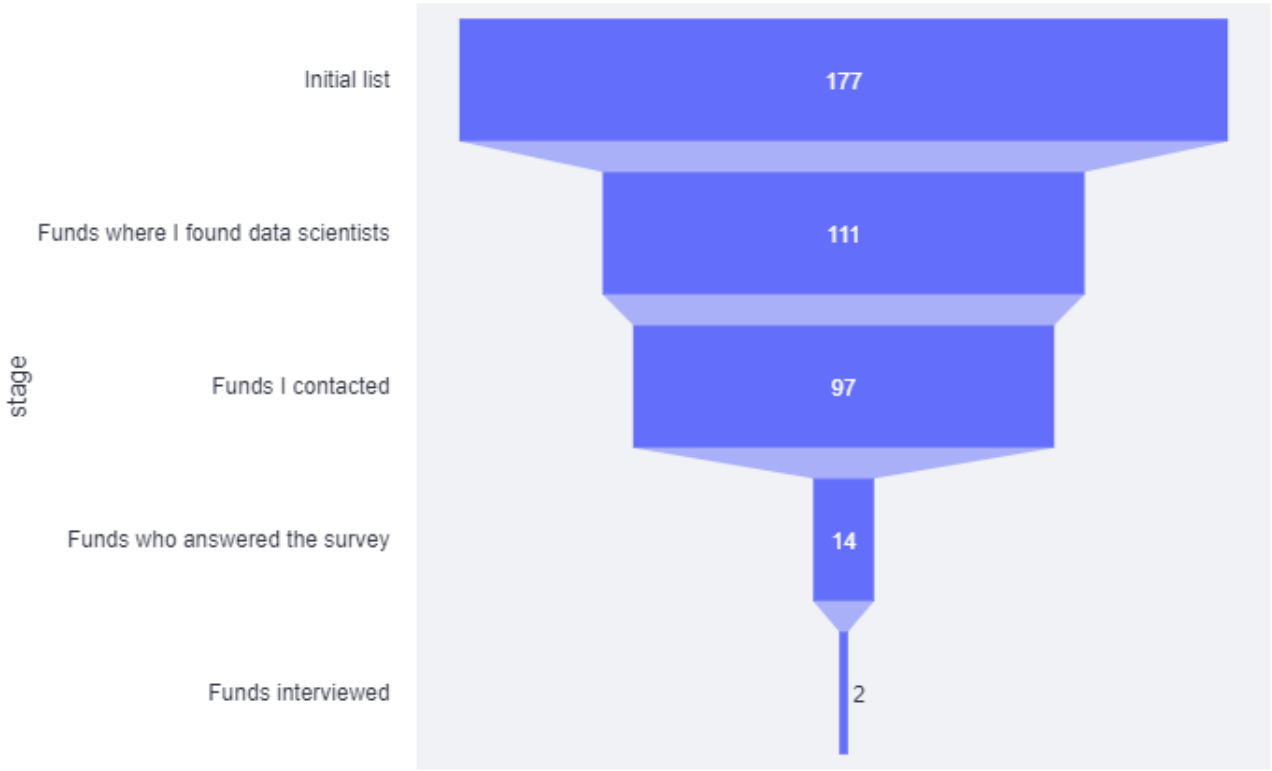


Figure 5: VC flowchart (Source : survey)

I therefore had a response rate of 14% on the answers to the survey, which I

find acceptable considering that I contacted people online.

5.2.2 Limitations and teachings

While carrying out the survey, I found some limitations in my work which could be corrected with more time on the survey.

- **Missing options** : while the multi-choice questions were never intended to be completely exhaustive in order to facilitate analysis, I noticed that my questions on the technologies used lacked some options. I focused too much on ML-orientated questions and forgot to include questions on intermediary technologies which already show an edge in analysis abilities. For instance, the options "SQL" and "BI tools (Tableau, PowerBI...) are not available by default. Thankfully, some respondents included those options in the free answer zone and we have at least partial data on this.

- **Rather small final sample** : even though the survey targeted very specialized individuals both mastering data science techniques and experience in Venture Capital, I only reached 14 answers at the time of the redaction of this Research Paper. While already informative, it is not enough to reach definitive conclusions during the analysis.

- Survivor bias : due to the nature of the survey, I may have only acquired data from funds which have no issues sharing information, even though the final data is anonymised. The initial search list I obtained is also affected by this : companies which manage to hide their data science usage may have a strong financial or technological incentive to do so.

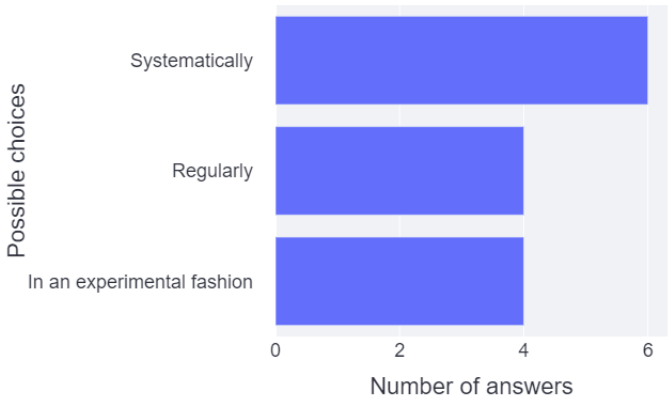
- **LinkedIn limitations** : in order to approach VC professionals in a more friendly way, I mostly avoided cold emails and instead turned to LinkedIn. However, LinkedIn limits messages and forced me in the free version to add people as relations, except when they had LinkedIn Premium and had configured their profile to accept InMail for people out of their network. While this was only slowing me down at the beginning, it almost spelled the end of the survey as I was detected by LinkedIn as having a Sales-related behavior : LinkedIn then prevented me from seeing any profile without a Premium account. I used my free trial

to switch to a Premium version including the Sales Navigator, which removed most roadblocks I had encountered. However, I still encountered random bugs where LinkedIn failed to acknowledge my Premium status, and was also surprised by the fact that the messaging system is not very ergonomic : any message received in the Sales Navigator first triggers a notification in the classical messaging system, which shows no new messages. You then have to move to the Sales Navigator tab to see the new message, which isn't good at all to track all sent messages.

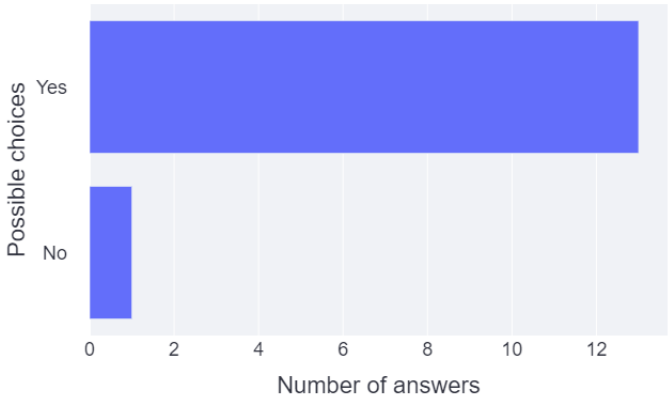
5.2.3 Results

Here are the raw results of the survey for each question.

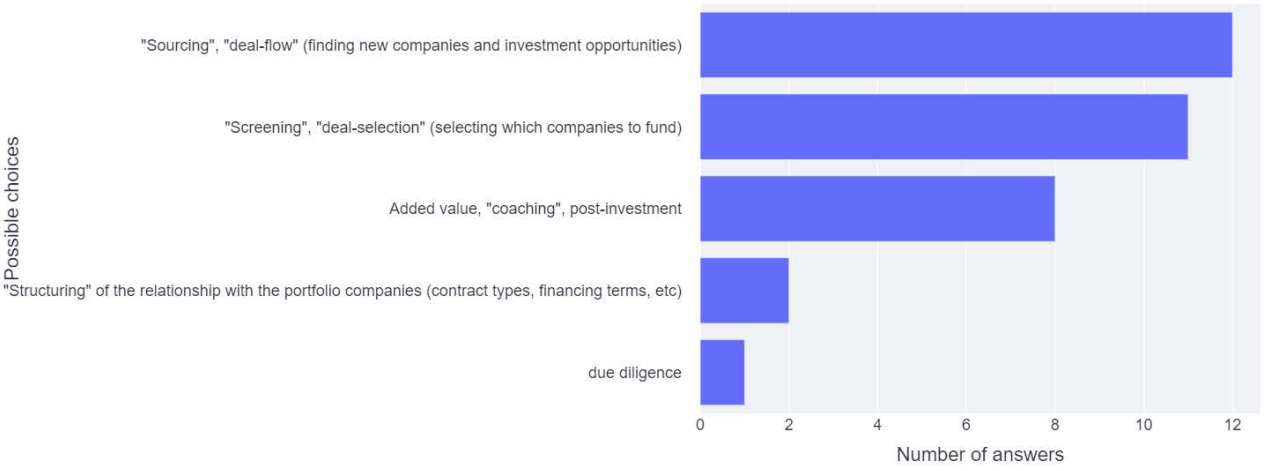
1) To which extent are you using data science techniques in your fund ?



2) Are you planning to deploy such technologies or to expand their use in the next 12 months ?

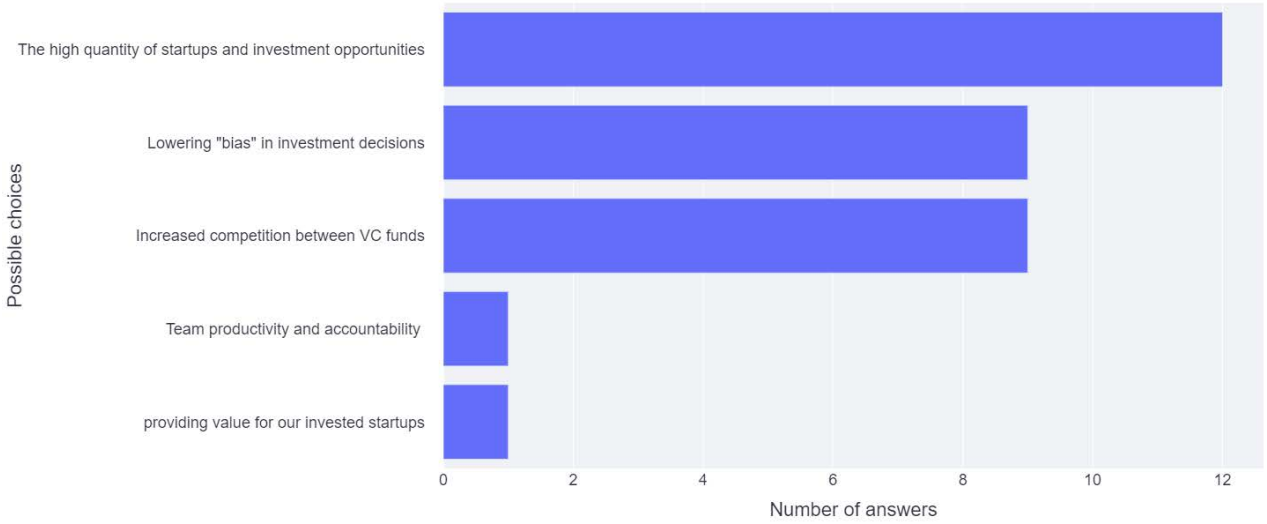


3) If you use such technologies, during which steps of the investing process do you use them ?

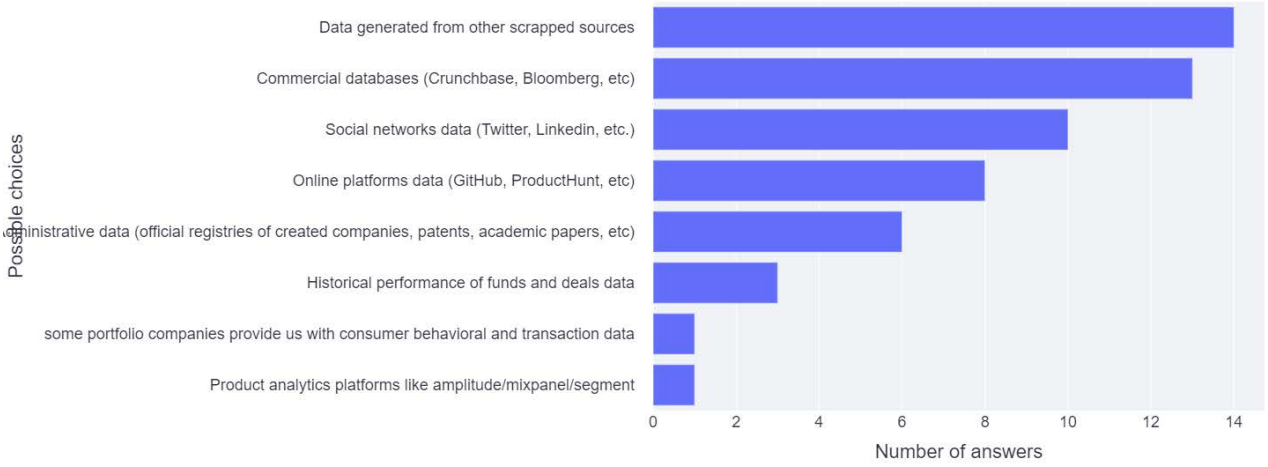


4)

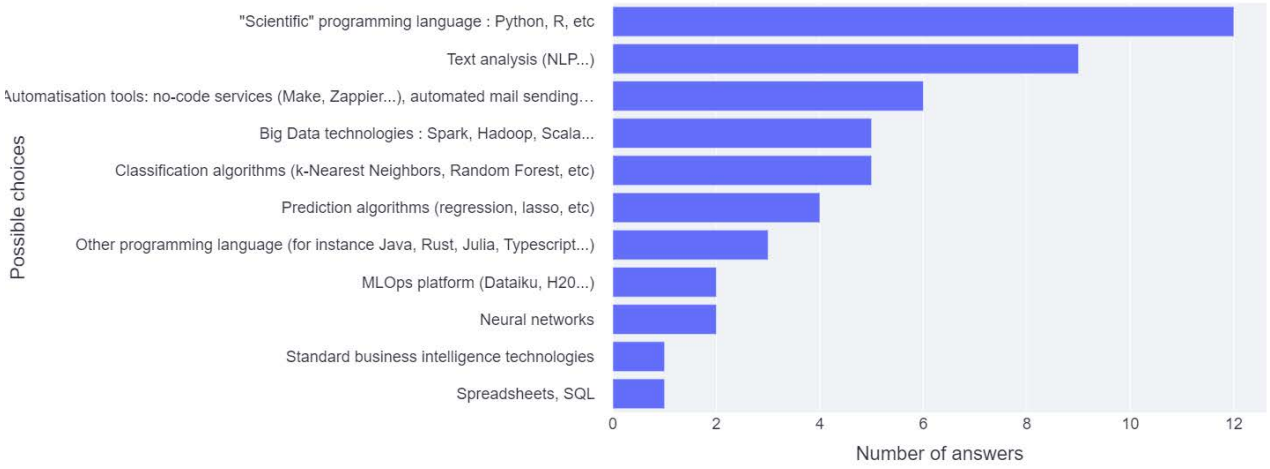
What are the main reasons driving your use of these technologies ?



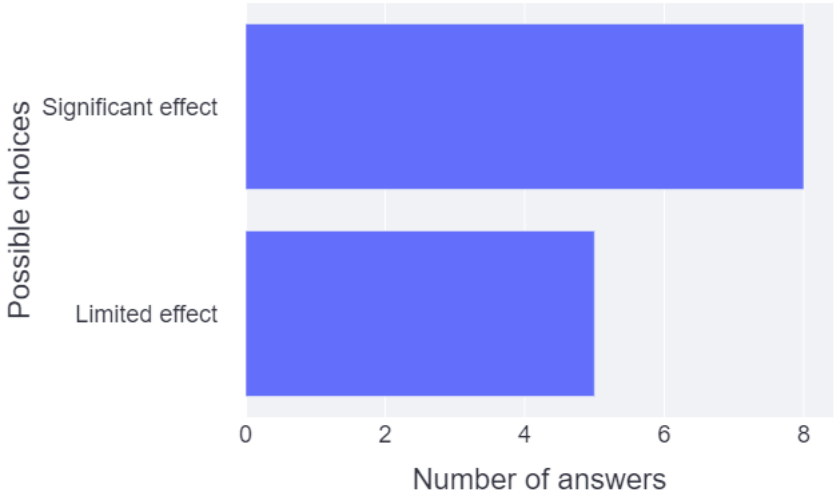
5) Which types of data sources are you using ?



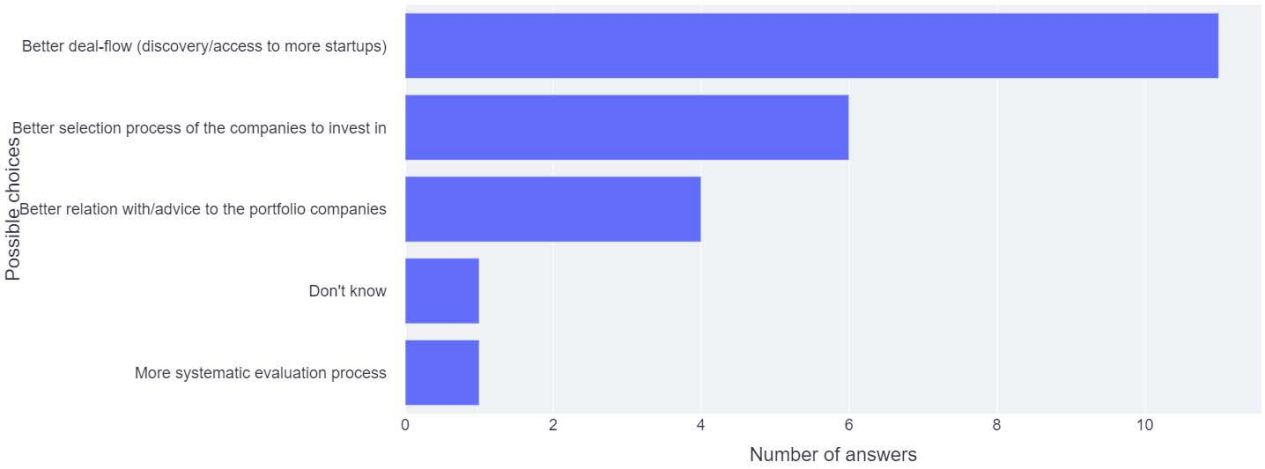
6) Which technologies are you using (algorithms types, programming languages/packages, on-board solution) ?



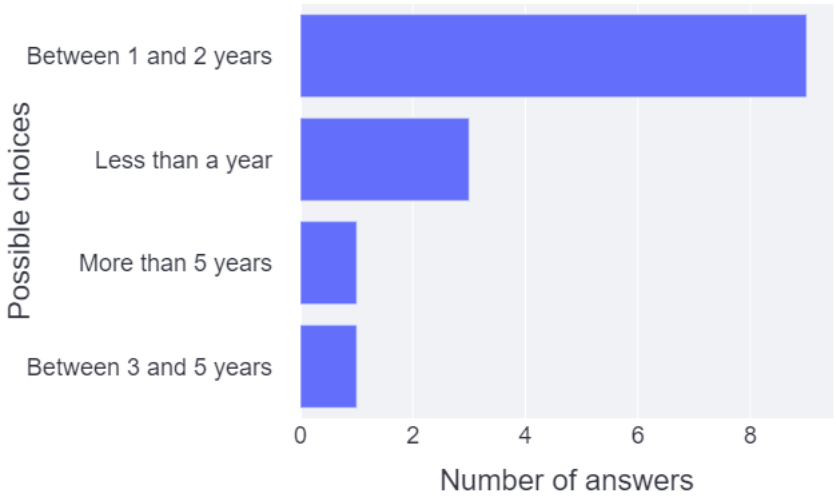
7) Do you consider that those technologies helped you to enhance the quality of your start-up investments ?



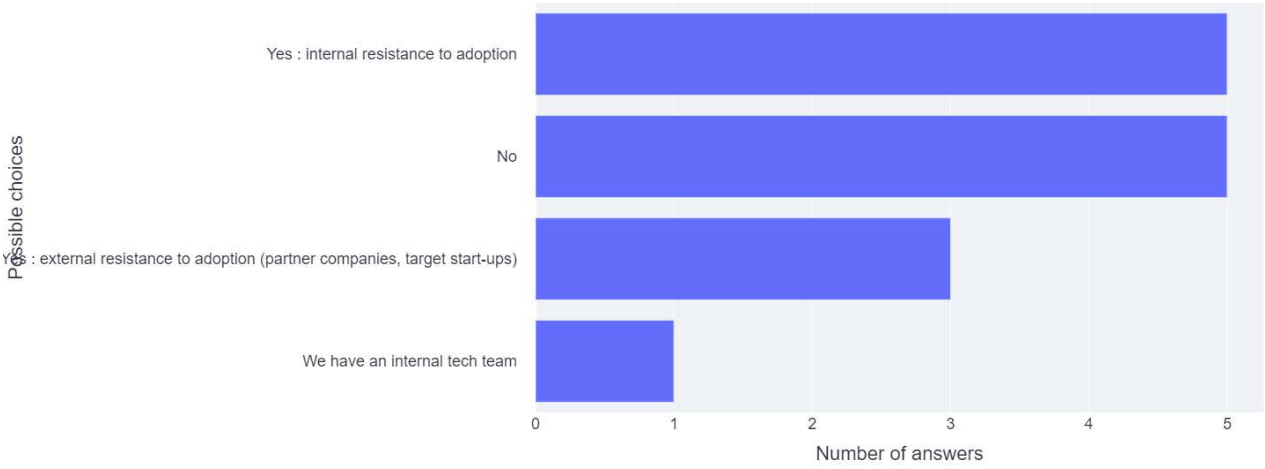
8) If they did, how ?



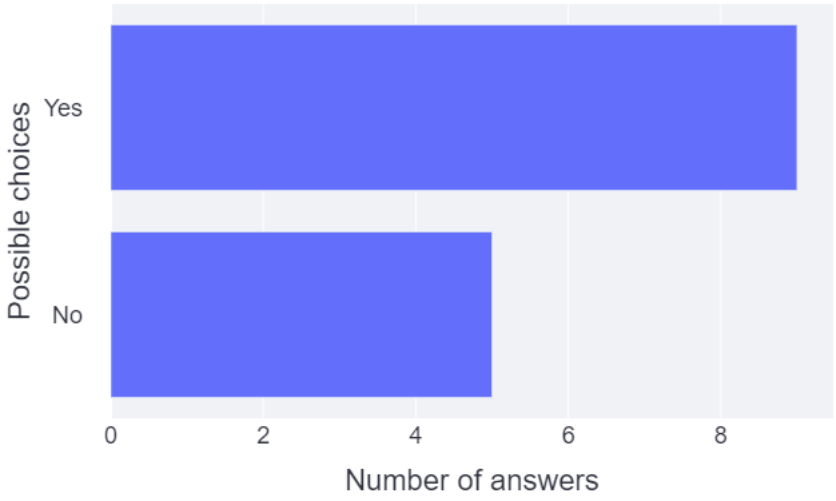
9) How long did it take to implement those solutions internally ?



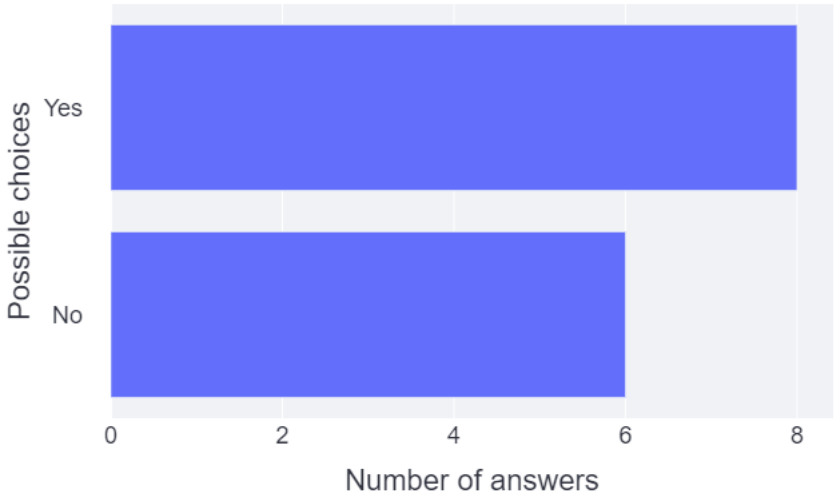
10) Do you encounter resistance to the implementation of data science techniques or to automatisation ?



11) Do you manage to find/recruit profiles who both master the business side and the technical side of data science implementation ?



12) Would you be willing to be contacted again for a brief interview in order to better understand the technologies that you are using ?



5.3 Analysis

5.3.1 Building analysis tools

In order to analyse the results of the survey, I designed a web-app available here : https://louistransfer-research-paper-analysis-home-co337u.streamlitapp.com/

It was built with the Streamlit python package, which both create a web server and a front-end from declarative python code. It is structured in 3 sections :

- A home page where the questions and the results of the survey can be explored;
- A uni-variate analysis page where individual data about the survey can be explored;
- A multi-variate analysis page where more complex analysis are performed.

Global architecture The webapp contains ingestion scripts which retrieve data from 2 sources : Google Forms and Google Sheets. The data is then automatically processed under the hood to generate the necessary datasets. Plots are generated with the Plotly package and can be exported as png files. On the front-end, the data is exportable in .xlsx and .csv formats. That way the results of the analysis can be replicated.

Retrieving Google Forms data First, the back-end automatically collects the latest survey data from the Google Form API, using the google-api-python-client package. I created a service account on Google Cloud Platform which handles the necessary authentication thanks to its client secret. Once the form raw response is received, it is processed in pandas and returns a cleaned dataframe with all answers to the survey. The emails and the timestamps are not loaded in order to anonymize the results.

Retrieving Google Sheets data Then, the service account fetches a Google Sheets containing the list of funds which have been contacted. It uses

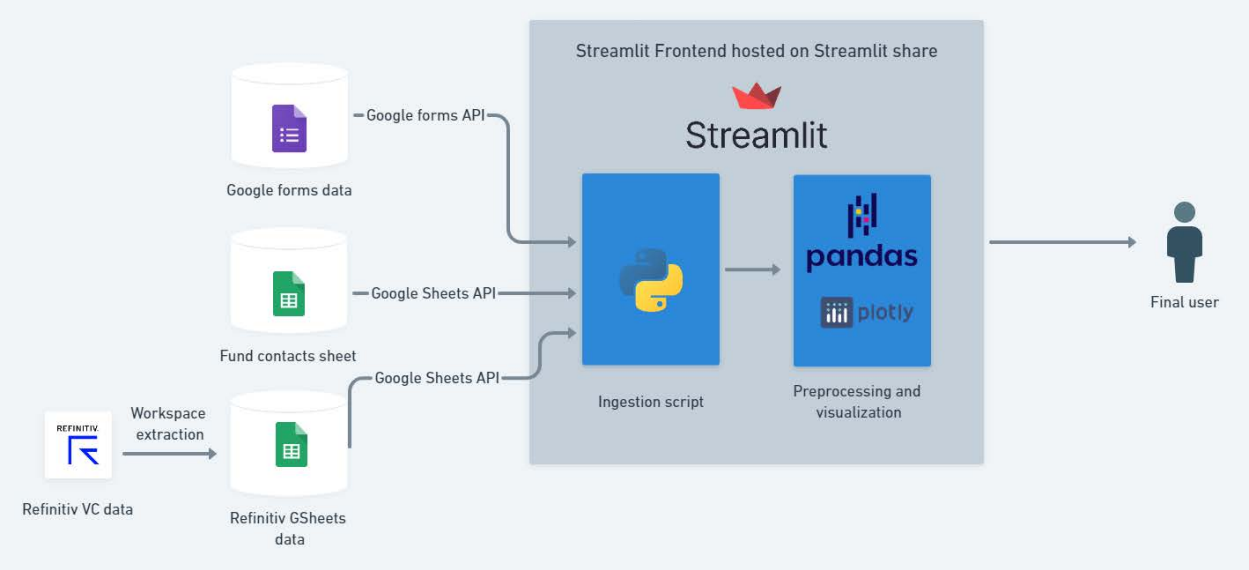


Figure 6: Architecture of the webapp used to automatically analyse the survey results

the same package and simply performs a slightly different authentication. It also imports data from Refinitiv extracts made in Refinitiv Workspace.

Experimenting with the Refinitiv API I wanted to import financial data about the funds which answered the survey as a way to detect correlations between answers and funds behavior. To do so, I requested a Refinitiv license from the HEC library and experimented on the Refinitiv Eikon API. The documentation of the API is very obscure, and the API itself is split in 3 products which have their own distinct python wrappers : eikon, refinitiv-data and refinitiv-data-platform.

Refinitiv only provides some code examples, and it even actively discourages developers from answering questions on a Stack Overflow model on their forums: developers must instead contact the Refinitiv Helpdesk. This severely limits the ability to find information on very specific issues. In particular, searching and filtering is not easy and requires the implementation of a "Screener" object, which corresponds to a list of objects in the Refinitiv

Workspace interface. One has to adapt Excel formulas in Python code. In my case, I was faced with an even more difficult issue : the Private Equity/VC Screener isn't compatible with the classical Screener interface used for instance to collect a list of M&A deals. Finally, I managed to get my hands on a custom script called "dataquery" released by a Refinitiv developer. [Sop] This script implements a very practical Screener class which simplifies the requets with the Refinitiv Eikon API and which manages to implement the Private Equity/VC screener.

However, since the PESCREENER is still in beta, most of the fields which are availble in the Workspace are in fact not available through the API. I then decided to perform classical extracts from the Workspace App, and to load them on Google Sheets in order to levarage the GSheets API in my Streamlit app.

5.3.2 Correlations with financial parameters

From Refinitiv Workspace, I gathered financial information about the performance of the 14 Venture Capital firms which had answered the survey, such as:

- The number of funds created during the lifetime of the VC Firm;
- The size of the sum of all funds of the firm (in million USD);
- The sum of all equity invested by the firm (in million USD);
- The number of companies that the firm invested in;
- The stage where the firm invested the most (ex: Early Stage)

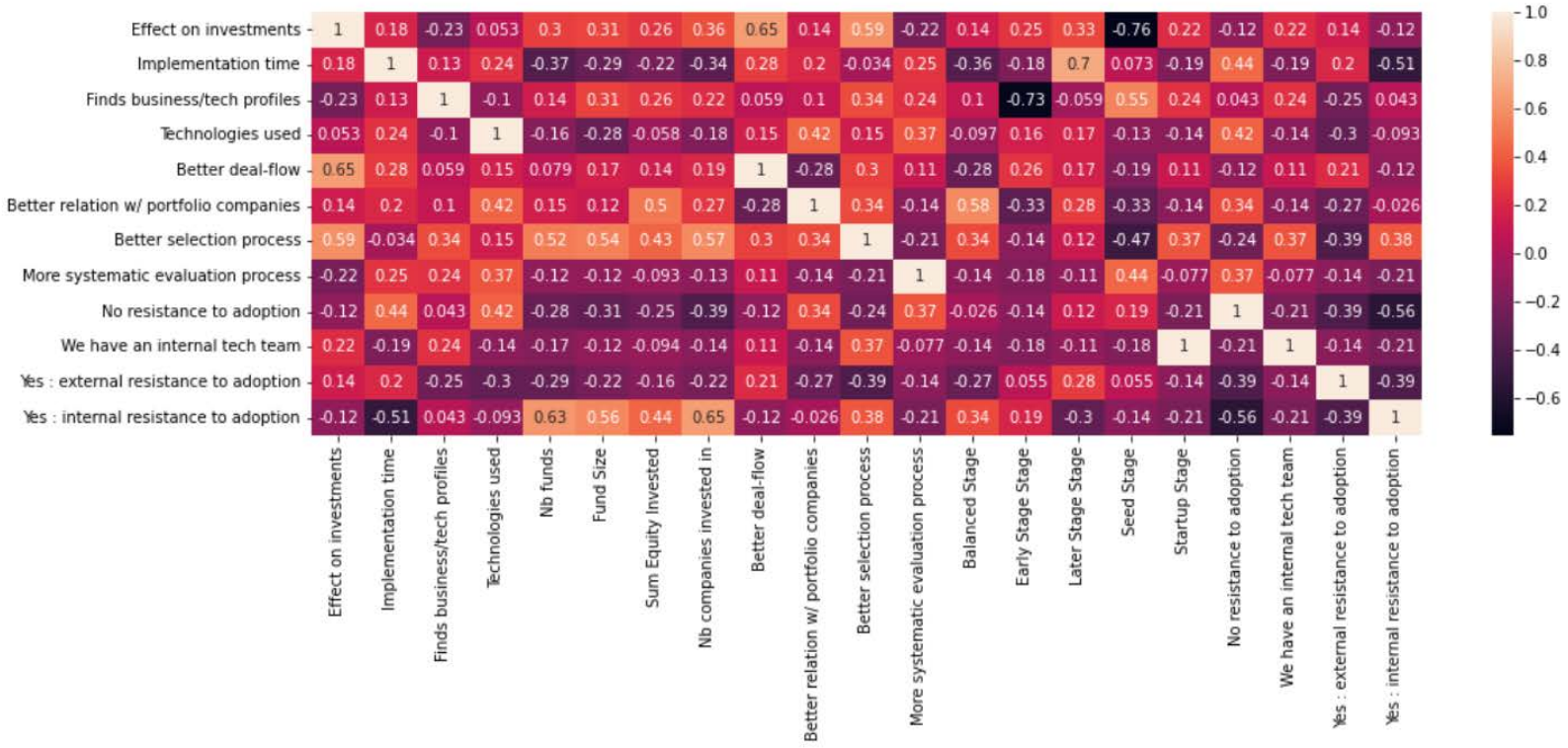
Using my survey tracking GSheets, I used the emails from Google Forms to join both datasets and to associate firms answers with their financial data. Finally, I converted the categorical variables from the survey into dummy variables, with the exception of the "Technologies used" column. In this case, I built a score to try to evaluate the data science maturity of the firm. This solution is of course very subjective, yet the final scores appear to reflect the level of "data maturity" of the firms. Here is the detail of the scoring (the

points are awarded when the answer has been selected in the multi-choice) :

Weights per answer	
Answer	Weight
"Scientific" programming language : Python, R, etc	1
Automatisation tools: no-code services (Make, Zapier...), automated mail sending...	1
Big Data technologies (Spark, Hadoop, Scala...)	3
Classification algorithms (k-Nearest Neighbors, Random Forest, etc)	3
MLOps platform (Dataiku, H20...)	3
Neural networks	5
Other programming language (for instance Java, Rust, Julia, Typescript...)	4
Prediction algorithms (regression, lasso, etc)	2
Spreadsheets, SQL	1
Text analysis (NLP...)	2

I used the final dataset to perform this correlations analysis. The matrix is rectangular on purpose : the rows only contain the variables produced by the survey in order to see how they interact with each other and with financial data.

Correlations



We can see two very interesting initial strong correlations. First, firms investing massively at the Seed stage also perceived the least effect of data science techniques on investments. This may be due to the fact that data science success metrics are not that clear during seed investment : is it the increased survival time of the chosen company or just a potentially blurry increased revenues due to data science ? Second, early stage firms have issues to find profiles who both handle the business and the technical side of data science.

The companies reporting a better selection process due to their use of data science experienced strong correlations with multiple variables. They tend to have experienced a stronger effect on investments, they usually have launched more funds than others and they are not seed companies. This tendency to use data for screening is reflected in the increased amount of companies they invest in and in the larger equity invested : screening is key when you have

to allocate more money in more companies on average as you have less time per investor to analyse each potential target.

I also found the strong correlation between the better deal-flow and the perceived effect on investments interesting. This means that either finding a very promising start-up first or freeing time for investors to dedicate themselves to other phases such as due diligence may be key to earn more per investment. This may also mean that the companies which have the ability to find those start-ups are in general more successful in their selection process.

The resistance to adoption is also interesting to analyse. It is inversely correlated with the technological score and with the implementation time, yet it is positively correlated with financial metrics showing a bigger organization size such as the fund size or the number of companies invested in. While the correlation is weak, Balanced stage firms experience the most internal resistance and Later Stage firms experience the most external resistance. This may show that the management of portfolio scale-ups is more reticent to accept data science projects with late stage VC firms, while the management of big VC firms investing in different stages may be the limiting factor in the deployment of internal data science solutions.

6 Building a new VC investment model : mixing data science decision-taking & gut feelings

6.1 Firms typology

From my experience exploring VC firms and interacting with VC investors, I tried to derive a typology of the main kinds of VC firms using data science I encountered on LinkedIn.

- **"Experimentation-phase firm"**: those are usually seasoned VC firms having multiple existing rounds of investments, which developed an interest in testing data science innovations. They have a broad investment range, and usually have small teams of data science which also act as investors or principals. They are trying to prove to partners that data science techniques can be useful.

- **"Data product VC"**: this type of firm has developed a real internal data science product. They usually employ data engineers, software engineers and ML engineers to build a mature data infrastructure which is most of the time used in sourcing or in screening. By design, those firms mostly focus on seed or series A investment stages;

- **"Seasoned data VC firm"**: those firms have gathered enough experience to develop a large internal data team. They will often employ PhDs to find breakthroughs in the field. While close to the "Data product VC", the aim is to keep the analysis in-house. The organization of their teams is also quite different, as those funds usually didn't start on a data science-centred investment thesis.

Using a more qualitative approach, I tried to apply this segmentation to the VC funds which answered the survey. Using a K-Means method and the analysis dataset I built during the survey, I built a small unsupervised model accounting for two situations :

- If the number of companies invested in by the fund is considered;

- If this variable is not considered.

Unfortunately, this approach is too sensitive to the number of companies invested in, and any combination not using this variable didn't produce any meaningful results. However, with more answers to the survey, this approach may pay off in the long-term.

6.2 How to implement and develop a data science practice

From this topology, it is possible to evaluate the ability to either build a VC firm around data science tools or to build a data science practice in an existing firm from scratch. I am going to focus on the "Experimentation-phase firm" kind of VC firm. I had the opportunity to interview a VC investor and data scientist working in a such a firm, who is the only data scientist in his team.

The fund invests a lot in the Seed stage. The companies usually come with their data, such as CRM and platform analytics data (number of web page unique visits...).

A key aspect of the interview was the idea of how to lower resistance to change in the fund. Indeed, this data-driven approach is usually quite different from the network-centric approach of VC investors who rely a lot on word-of-mouth to discover new start-ups. To do so, the data scientist I interviewed had to carry out the message that those technologies do not replace the existing network of contacts of the fund, they only help to add additional data. Education is key to share why the platform is useful. Since metrics are difficult to come by in the Seed phase, performing post-mortems is useful to show that a successful company probably wouldn't have been detected without data science.

There are two ways to build a data platform at this point : to completely outsource the platform, or to recruit data engineers, software engineers and data engineers internally to build a data team. In his fund, the platform has been developed by free-lancers, he therefore acts as a data product manager who specifies the pain points the platform will have to address. Recruiting in data science in Europe is more recent than in the US, as most innovations in

data science have been made by "Data Product VC" firms in San Francisco. Therefore, a company acculturation phase with a small team is certainly necessary before recruiting more data science profiles.

Another interesting practice to start is data science operations : another fund also evoked this new discipline on the phone. The idea is to provide data science tools to portfolio companies for them to get an advantage over the markets. For instance, CRM segmentation tools or the ability to start an AWS or GCP cloud instance on the firm's account with a seasoned data scientist can prove invaluable in the first months of the start-up.

6.3 Conclusion and prospective

With this analysis, I have been able to describe the data science field in Venture Capital, even though the number of answers to the survey isn't that large for the moment. Those techniques can work for multiple stages of the life of VC funds portfolio companies, and they help to reduce time spent on sourcing and checking on companies. This may free more time for VC investors to coach their portfolio companies, which may in turn increase productivity in those companies.

It appears from the results of the analysis that data science techniques are bound to grow even further, especially in Early Stage. Even though the rise of Central Banks rates has started to limit the supply of cash available and may reduce the pool of start-ups to invest in, a VC investor told me during a mail exchange that it wouldn't impact their use of their screening tool. Indeed, the time advantage granted by this tool is still viable even in this kind of economic context.

In the coming years, it will be really interesting to see if the phenomenon of the "Lucas Critique" develops in this sector : as more and more funds gain access to the knowledge offered by advanced data science techniques such as automated scraping providers, it will be more and more difficult to be the first to find the next unicorn. Start-up founders may then have to adapt their Linkedin or GitHub profiles quickly in order to attract the attention of algorithms, while any change on the market will have to be quickly translated in the ML algorihms used by data sicence professionals in Venture Capital firms.

7 Bibliography

References

[Mil11] Claire Cain Miller. "Google Looks for the Next Google". In: *The New York Times* (2011). URL: https://www.nytimes.com/2011/07/20/technology/google-spending-millions-to-find-the-next-google.html.

[Kap17] Nikhil Kapur. *How a VC funnel works*. 2017. URL: https://grayscale.vc/how-a-vc-funnel-works-6f1202d0ac9.

[Neu19] Jerry Neuman. *Why do VCs insist on only investing in high-risk, high-return companies?* 2019. URL: http://reactionwheel.net/2019/01/why-do-vcs-insist-on-only-investing-in-high-risk-high-return-companies.html.

[CBC21] Francesco Corea, Giorgio Bertinetti, and Enrico Maria Cervellati. "Hacking the venture industry: An Early-stage Startups Investment framework for data-driven investors". In: *Machine Learning with Applications* 5 (2021), p. 100062. ISSN: 2666-8270. DOI: https://doi.org/10.1016/j.mlwa.2021.100062. URL: https://www.sciencedirect.com/science/article/pii/S2666827021000311.

[Blo+22] Ivo Blohm et al. "It's a Peoples Game, Isn't It?! A Comparison Between the Investment Returns of Business Angels and Machine Learning Algorithms". In: *Entrepreneurship Theory and Practice* 46.4 (2022), pp. 1054–1091. DOI: 10.1177/1042258720945206. eprint: https://doi.org/10.1177/1042258720945206. URL: https://doi.org/10.1177/1042258720945206.

[RME22] Nathan Reiff, Julius Mansa, and Ryan Eichler. "Series A, B, C Funding: How It Works". In: *https://www.investopedia.com/articles/personal-finance/102015/series-b-c-funding-what-it-all-means-and-how-it-works.asp* (2022).

[For] Venture Forward. *Venture History 101*. URL: https://ventureforward.org/education/history-101/#vcservicemodel.

[NG] Mathieu Nasri and Ieva Gaigala. "Les traits neurologiques des meilleurs VCs : une étude inédite sur des facteurs de succès jusque-là insoupconnés". In: (). URL: https://docsend.com/view/h4tji5f6q2nctykn.

[Sop] Leonid Sopotnitskiy. *Dataquery script*. URL: https://github.com/Refinitiv-API-Samples/dataquery.

A Appendix

A.1 Other plot analyses

Classification of firms which answered the survey represented as a function of the number of companies they invested in and their perceived level of technological advancement

Figure 7: This plot was produced with a K-Means algorithm, in order to find unsupervised groups of VC firms depending on their characteristics. However, the number of companies invested is the only strong predictor, and the level of technological advancement isn't enough to segment new intra-groups which would have been interesting to analyse.

A.2 Refinitiv API code used

Here is some Python code I wrote to leverage the Screener object from the dataquery.py script when I was experimenting with the Refinitiv Eikon API :

```
# Set up
import eikon as ek
import refinitiv.data as rd
from refinitiv.data.content import search
from dataquery import screener, FORMULA, PRIVATEEQUITY
APP_KEY = <Refinitiv App key here>
ek.set_app_key(APP_KEY)
rd.open_session()
```

```
def get_deals_list(fund_name, fields):
    sc = screener(
        FORMULA(f'Contains(TR.PEInvestFundInvestorName,
        ↪  "{fund_name}")'),
        universe=PRIVATEEQUITY,
        currency="USD",
    )
    query = sc.query
    df, err = ek.get_data(
        query,
        fields = fields,
    )
    if "TR.PEInvestCompanyAllInvestorFunds" in fields:
        df["Investee Company All Investor Funds"] =
        ↪  df["Investee Company All Investor
        ↪  Funds"].str.split(';')
    return df

FIELDS = [
            "TR.PEInvestCompanyName",
            "TR.PEInvestFundInvestorName",
            "TR.PEInvestCompanyNation",
            "TR.PEInvestCompanyPrimarySIC",
            "TR.PEInvestRankValue",
            "TR.PEInvestCompanyEstEquityReceivedToDate",
            "TR.PEInvestCompanyFirstInvestmentReceivedDate",
            "TR.PEInvestCompanyLastInvestmentReceivedDate",
            "TR.PEInvestCompanyPortfolioStatus",
            "TR.PEInvestCompanyStatus",
            "TR.PEInvestCompanyAllInvestorFunds"
```

```
        ]
df = get_deals_list(<VC_fund_name>, fields=FIELDS)
df
```

And here is the code used to assemble various extracted datasets from the Refinitiv Workspace software :

```
COLS_TO_KEEP = [
        "Firm Name",
        "Nb funds",
        "Fund Size",
        "Firm Stage",
        "Firm Industry",
        "Sum Equity Invested",
        "Estimated Equity Available",
        "Nb companies invested in",
        "Sum Deal Rank"]

col_replace_dict = {
    "No. of Funds": "Nb funds",
    "Firm's Fund: Size\n(USD, Millions)": "Fund Size",
    "Total Known Equity Invested by Fund\n(USD, Millions)":
    ↪  "Sum Equity Invested",
    "Estimated Equity Available\n(USD, Millions)": "Estimated
    ↪  Equity Available",
    "Total Number of Companies Invested in by Fund": "Nb
    ↪  companies invested in",
    "Deal Rank Value\n(USD, Millions)": "Sum Deal Rank",
    "Firm Investors Stage Investment Preference\n('|')": "Firm
    ↪  Stage",
    "Firm Investors Industry Investment Preference\n('|')":
    ↪  "Firm Industry"
}

def generate_full_refinitiv_dataset(refinitiv_investments_path:
↪  str, refinitiv_fundraising_path: str, cols_to_keep: list):

    df_refinitiv_investments =
    ↪  pd.read_excel(refinitiv_investments_path)
```

```
    df_refinitiv_investments = df_refinitiv_investments.rename(
        columns={"Firm Investor Name": "Firm Name"}
    )
    df_refinitiv_investments = df_refinitiv_investments[
        df_refinitiv_investments["Firm Name"] != "TOTAL"
    ]
    # df_refinitiv_investments

    df_refinitiv_fundraising =
    ↪ pd.read_excel(refinitiv_fundraising_path)
    df_refinitiv_fundraising = df_refinitiv_fundraising[
        df_refinitiv_fundraising["Firm Name"] != "Industry
        ↪ total"
    ]

    df_refinitiv_full = df_refinitiv_fundraising.merge(
        df_refinitiv_investments, on="Firm Name", how="left"
    )

    df_refinitiv_full =
    ↪ df_refinitiv_full.rename(columns=col_replace_dict)

    df_refinitiv_full = df_refinitiv_full[cols_to_keep]

    targets_list_data =
    ↪ pd.unique(df_forms_completed["Fonds"]).tolist()
    df_refinitiv_full["Main Fund Name"] =
    ↪ df_refinitiv_full["Firm Name"].apply(
        lambda sub_fund_name: search_fund_name(sub_fund_name,
        ↪ targets_list_data)
    )
    return df_refinitiv_full

def refinitiv_data_cleaning(df_refinitiv_full):
    if df_refinitiv_full["Firm Name"].isin(["OC4 Ventures Fund
    ↪ I LP"]).any():
```

```
        df_refinitiv_full =
        ↪  df_refinitiv_full[df_refinitiv_full["Firm
        ↪  Name"]!="OC4 Ventures Fund I
        ↪  LP"].reset_index(drop=True)
    df_refinitiv_full.loc[df_refinitiv_full["Firm
    ↪  Industry"].isna(), "Firm Industry"] = "Diversified"
    df_refinitiv_full.loc[df_refinitiv_full["Firm
    ↪  Stage"].isna(), "Firm Stage"] = "Balanced Stage"
    return df_refinitiv_full
```

A.3 Streamlit app code example

Here is some code I built to create the webapp used to generate the plots :

```
import os
import io
import toml
import streamlit as st
import pandas as pd
from logzero import logger
from pathlib import Path
from analyser.api_helpers import (
    authenticate,
    import_form_data,
    import_gsheet_data,
    open_yaml,
)
from analyser.streamlit_helpers import check_password
from analyser.viz_helpers import generate_funnel

CONFIG_PATH = os.path.join(".streamlit", "secrets.toml")

config = toml.load(CONFIG_PATH)

SCOPES = st.secrets["api"]["default_scopes"]
SERVICE_ACCOUNT_INFO = st.secrets["api"]["service_account"]

DATA_PATH = st.secrets["data"]["default_data_path"]
FORMS_CSV_PATH = os.path.join(DATA_PATH, "full_forms_data.csv")
```

```
GSHEETS_CSV_PATH = os.path.join(DATA_PATH, "gsheets_data.csv")
QUESTIONS_YAML = st.secrets["data"]["yaml_questions_path"]
TRANSLATION_YAML = st.secrets["data"]["yaml_translation_path"]
FRENCH_FORM_ID = st.secrets["forms"]["french_form_id"]
ENGLISH_FORM_ID = st.secrets["forms"]["english_form_id"]
GSHEET_ID = st.secrets["sheets"]["gsheet_id"]
MULTI_CHOICE_COLUMNS =
↪  st.secrets["forms"]["multi_choice_columns"]
RAW_NEW_COLUMN_NAMES = st.secrets["forms"]["new_column_names"]
NEW_COLUMN_NAMES = {int(key): value for key, value in
↪  RAW_NEW_COLUMN_NAMES.items()}
INTRODUCTION_PATH = st.secrets["forms"]["introduction_path"]

# Initial loading

if check_password():

    questions_mapping = open_yaml(QUESTIONS_YAML)
    translation_mapping = open_yaml(TRANSLATION_YAML)
    api_forms = None
    api_sheets = None
    if not os.path.exists("data"):
        os.mkdir("data")

    if os.path.exists(FORMS_CSV_PATH):
        logger.info("Found existing csv, loading it now.")
        converter_dict = {i: pd.eval for i in
        ↪  MULTI_CHOICE_COLUMNS}
        df_full = pd.read_csv(FORMS_CSV_PATH, sep=";",
        ↪  converters=converter_dict)
        df_full.columns = ["create_time", "email"] +
        ↪  list(range(2, 14))
        # df_full = df_full.drop(columns=['email'])
    else:
        logger.warning("No csv found, extracting data from the
        ↪  Forms API.")
        api_forms, api_sheets =
        ↪  authenticate(SERVICE_ACCOUNT_INFO, SCOPES)
        df_full = import_form_data(
```

```
            api_forms,
            FRENCH_FORM_ID,
            ENGLISH_FORM_ID,
            questions_mapping,
            translation_mapping,
            FORMS_CSV_PATH,
        )

    if os.path.exists(GSHEETS_CSV_PATH):
        logger.info("Found existing csv, loading it now.")
        df_gsheet = pd.read_csv(GSHEETS_CSV_PATH, sep=";")
    else:
        logger.warning("No csv found, extracting data from the GSheets API.")
        api_forms, api_sheets = authenticate(SERVICE_ACCOUNT_INFO, SCOPES)
        df_gsheet = import_gsheet_data(
            api_sheets, GSHEET_ID, GSHEETS_CSV_PATH, range="Contacts"
        )

    st.title("Research paper analysis")

    col1, col2 = st.columns(2)

    with col1:
        st.metric("Number of answers to the survey ", df_full.shape[0])

    with col2:
        if st.button("Reload GForm data"):
            if api_forms == None:
                api_forms, _ = authenticate(SERVICE_ACCOUNT_INFO, SCOPES)
            df_full = import_form_data(
                api_forms,
                FRENCH_FORM_ID,
                ENGLISH_FORM_ID,
                questions_mapping,
```

```
92                 translation_mapping,
93                 FORMS_CSV_PATH,
94             )
95
96         if st.button("eload GSheets data"):
97             if api_sheets == None:
98                 _, api_sheets =
                   ↪  authenticate(SERVICE_ACCOUNT_INFO, SCOPES)
99             df_gsheet = import_gsheet_data(
100                api_sheets, GSHEET_ID, GSHEETS_CSV_PATH,
                   ↪  range="Contacts"
101            )
102
103    with st.expander("Working paper description"):
104        st.markdown(Path(INTRODUCTION_PATH).read_text())
105
106    options = df_full["email"].unique()
107    selected_options = st.multiselect("Choose emails to filter
       ↪  the dataframe", options)
108    # selected_options = options
109    if selected_options:
110        current_data =
           ↪  df_full[df_full["email"].isin(selected_options)]
111    else:
112        current_data = df_full
113    current_data =
       ↪  current_data.rename(columns=NEW_COLUMN_NAMES)
114    st.dataframe(data=current_data, width=1200)
115
116    col3, col4 = st.columns(2)
117
118    with col3:
119
120        st.download_button(
121            label="Download data as CSV",
122            data=current_data.to_csv(encoding="utf-8", sep=";",
                   ↪  index=False),
123            file_name="survey_results.csv",
124            mime="text/csv",
```

```
        )

        # Write each dataframe to a different worksheet.
    with col4:
        buffer = io.BytesIO()
        with pd.ExcelWriter(buffer, engine="xlsxwriter") as
        ↪  writer:
            current_data.to_excel(writer)
            st.download_button(
                label="Download data as Excel",
                data=buffer,
                file_name="survey_results.xlsx",
                mime="application/vnd.ms-excel",
            )

    st.title("Answers funnel")

    fig = generate_funnel(df_gsheet, df_full)

    st.plotly_chart(fig)
```