Using data science to enhance Venture Capital deals

Overview of the data science field in the Venture Capital industry

Why should we evaluate the use of data science in the VC industry ?

Rising trend in finance

> Algorithmic trading and use of **massive datasets** by hedge funds

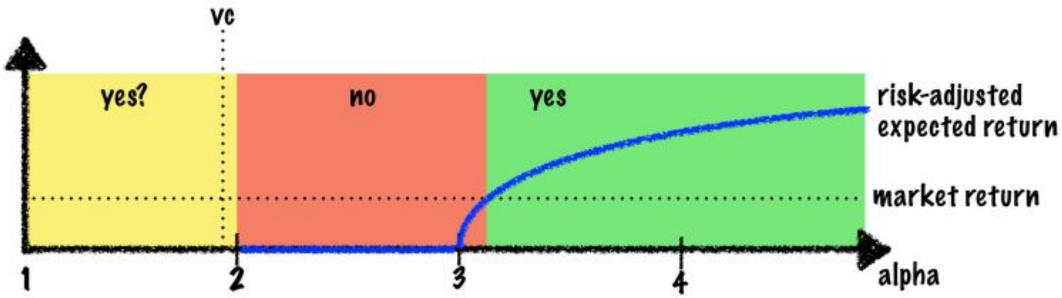
> ML models used in production in banks to create **new customized services** for clients

Very specific environment

> Investments are **riskier by design** in the VC industry

> Yet heuristics are everywhere in VC investment : past experience is **replicated on new deals** even if the situation differs

> Biases affect investors who tend to **rely on their network** to discover start-ups



Objectives and hypotheses

Study objective

> Understanding the **maturity** of data science practices in VC funds

> Evaluating whether some investment stages and some parts of the deal pipeline are **more targeted by those technologies** than others

> Determining **why** those funds use data science

> Figuring out **how to start a data science practice** in an existing fund

Hypotheses

> Judging the current techniques put forward in academic papers, most techniques should be deployed during the sourcing/screening phases of the deal pipeline

> The funds which use them should have a certain critical size in order to pay for those technologies

Gathering information on data science usage

Methodology

> Targeting **highly specialized individuals** working in data science-related jobs in Venture Capital Firms

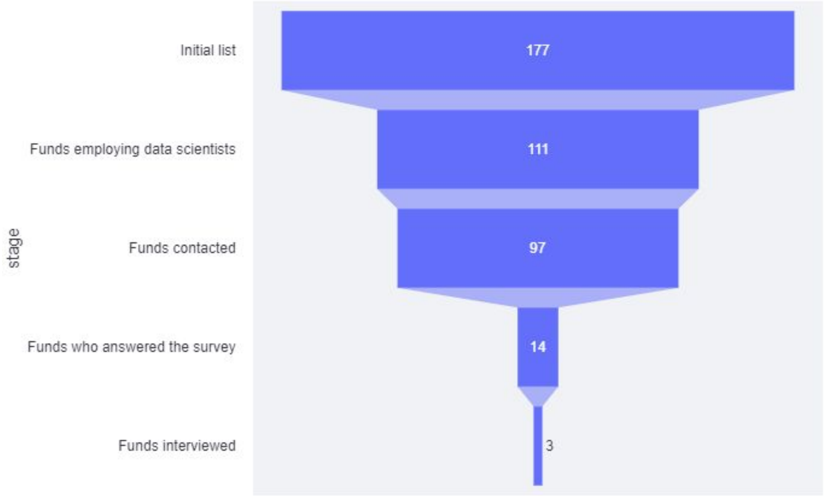
> Using LinkedIn to **select profiles** and to contact them -> reducing the difficulty of reaching out through cold emails

> I used **LinkedIn Premium** since I was detected by LinkedIn as checking on too many profiles

Results

> 14 final answers

> Three interviews



Results

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

> 6 firms are using it systematically, 4 regularly and 4 experimentally

> The deployment of data science solutions is **quite mature** on the market

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

> 13 yes, 1 no

> **Overwhelming intention** to expand those technologies in the future

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

> 12 use it during sourcing, 11 during screening, 8 during coaching, 2 during the structuring of the relationship and one during the due diligence

> It confirms that **sourcing and screening are the key targets**. Yet one surprise, **coaching is also becoming quite mainstream** -> might be worth investigating further (it confirms information from two interviews I had)

Results

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

> 12 firms use it because there is a high quantity of startups and investment opportunities, 9 use it to lower bias in investment decisions, 9 because of increased competition between VC funds, one use it for team productivity and accountability and one uses it to provide value for invested start-ups

> Two main effects are visible here : these technologies are mainly used to **increase the quantity and the quality of investments**. The idea is to select **more start-ups** with **a lower investment bias** in order to compensate the investors potential flaws.

Which types of data sources are you using ?

> 14 use data generated from other scraped sources, 13 use commercial databases, 10 use social networks data, 8 use online platforms data, 6 use administrative data, 3 use historical performance of funds, 1 use consumer behavioral data from funds and one uses product analysis platforms

> **Further interviews** would be interesting to determine what the “other scraped sources” are. Furthermore, **historical data of funds** is surprisingly **not used that much**.

Results

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

> 6 use scientific programming language, 3 use big data tech, 1 use other programming languages, none use MLOps platform, 2 use automation tools, 4 use classification algorithms, 3 use prediction algorithms, 6 use text analysis, one uses neural networks, one uses spreadsheets/SQL, one uses standard BI tools

> The key technologies used by the funds are Python and NLP text analysis. The use of ML algorithms show that the funds probably **parse texts from scrapped sources** in order to **evaluate the capabilities of a start-up** (technology, team experience...)

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

> 8 firms have seen significant effect from the use of this technology, 5 have seen limited effects and no funds report an absence of effect

> Data scientists report an **increase in the quality of the investments** from the use of data science. The result may be **biased** here though because funds which failed to implement a practice do not appear in the survey

Results

If they did, how ?

> 5 firms reported a better deal-flow with the discovery of more startups, 3 reported a better selection process of the companies to invest in, 3 reported a better relationship with portfolio companies, one reported a more systematic evaluation process and one doesn’t know.

> This once again confirms that **sourcing**, **screening** and **coaching** are the steps where **data science methods are the most effective**.

How long did it take to implement those solutions internally ?

> 9 firms took between one and 2 years to implement their techniques, 3 took less than a year, one took 3 to 5 years and one took more than 5 years

> The analysis of this question could be improved with a **higher response rate to the survey** : with more answers, we could evaluate whether firms which took more than 2 years are **more effective** than the ones which took one to 2 years.

Results

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

> 5 firms do not experience resistance to adoption, 5 firms see internal resistance to adoption and 3 firms experience external resistance to adoption

> Surprisingly, there is **quite a lot of external resistance to adoption**.

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

> 9 firms do manage to find such profiles, while 5 are struggling to find them

> I was surprised by this answer, which may indicate that firms are **attractive enough** to find candidates which are usually hard to find in other industries.

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

> 8 people can be contacted for an interview, 5 are not interested.

Behavioral science

Discovering Unfair

> I listened to a podcast with Daphni evoking Unfair : https://open.spotify.com/episode/3eQlBWh6I7N7tKV58aSdo1?si=3002f6cf812a4951

> I decided to contact the CEO, he described the **technology** he sells to VC firms

How it works

> The co-founders of a company are tested with questions which **determine their neurological profile**, then the software evaluates their success rate

> Same for investors in order to **reduce investment biases**

Why it matters

> Solution used by a **lot of French VC firms**

> Built **hand-in-hand with investors** from VC firms

> **89%** success rate on a post-mortem analysis

> Further work on behavior may be interesting : **lots of behavioral scientists in US VC firms**

Biases and the importance of funds' size

Effect of funds' size

> A potential bias exists : big funds in size will have the money to use these technologies

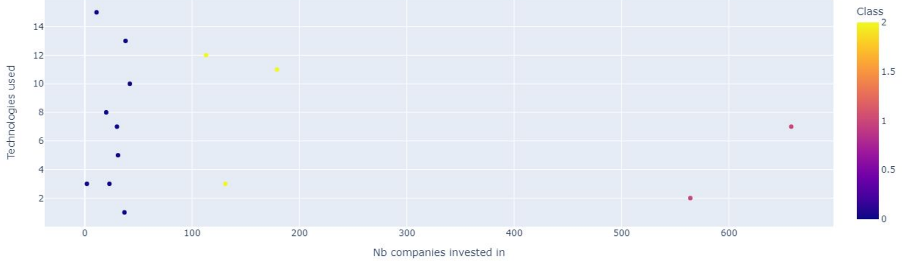
> Paradox : the companies which do not need the technologies have them, while smaller companies which need it do not have the money to invest in this kind of tech

> While mid-sized funds appear to be a good compromise, current economic situation tends to force either a conversion into a small specialized fund or a growth into a big fund

Other experience of this "size effect"

> I tried to segment funds into categories with a K-Means classification algorithm

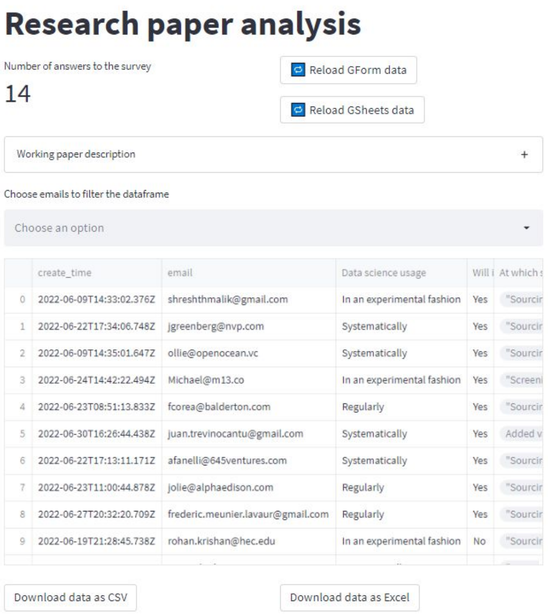
> Only one feature "fixed" the classification once and for all : the number of companies invested in by the fund. Technological maturity didn't affect the result or any other answers to the survey

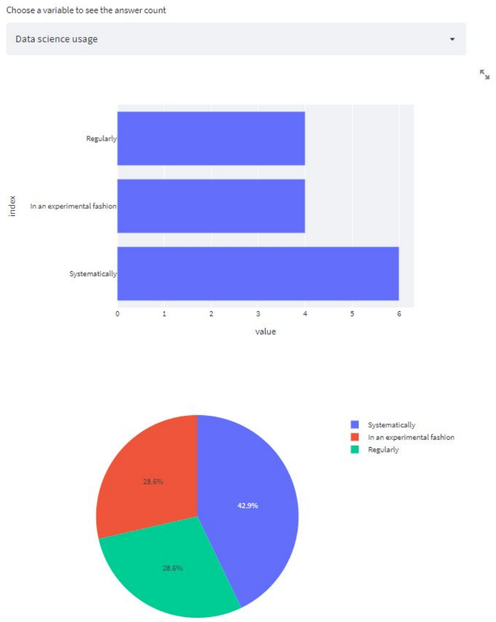


Streamlit webapp

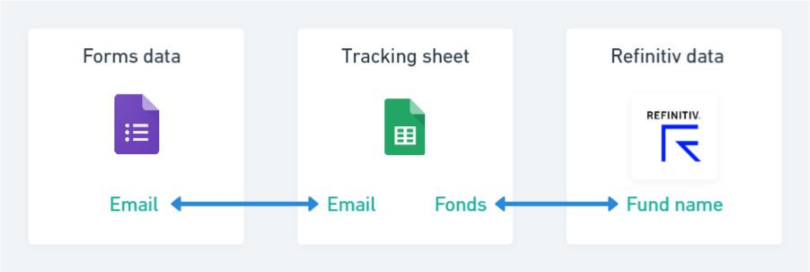
> https://louistransfer-research-paper-analysis-home-co337u.streamlitapp.com/

> Some modifications must be made to the app (including Refinitiv data and correlation plots)

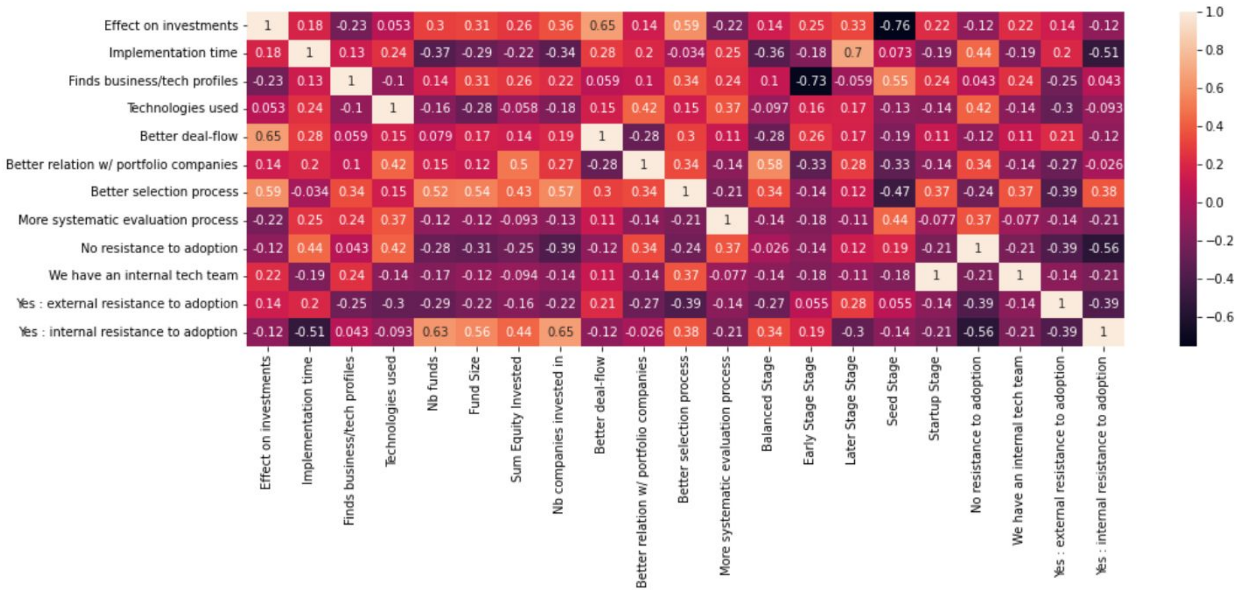




Joining the data



Correlation analysis



Further analysis and survey improvements with more time

01	**Contacting more data science professionals**	• Increase the sample size • Will show more types of VC firms
02	**Improving the streamlit app**	• Including the last Refinitiv analyses • Prototype in the work, one more day of work should be enough
03	**Improved clustering**	• Once the sample is improved, trying to run my clustering again (K-Means) • Redefining a typology once it is done
04	**Building a start-up selection tool**	• Using the Crystal LinkedIn app in an automated fashion • Using Refinitv data automatically to try to find promising start-ups
05	**Improving ties with professionals in the sector**	• Being able to track changes in the long term

Conclusion

State of the market

> In 2022, data science is **truly present in VC firms.**

> It is mainly used in **sourcing**, **screening** and **coaching**. "**Data Ops**" seems to be a new emerging trend for VC firms as a way of coaching.

> Some funds build **complete products** which can be sold to other companies. They can either be outsourced or built internally. The team will then have software engineers and data engineers in its LinkedIn employees.

> Others will instead **experiment** with a much smaller team in order to **prove the efficiency** of data science methods.

Answers to the hypotheses

> **Sourcing and screening** are indeed key. However, evaluating the effect of **coaching** in further academic papers may make sense to see if the coached portfolio companies perform better than their non-coached competitors on their market.

> There is a **true size effect** on the use of data science, which may **cloud the effect of technologies on the efficiency of a fund**. Increasing the survey sample size is necessary to try to bypass this. **Contacting former VC firms data scientists** may also help to reduce the survivor bias.

Conclusion

Prospective

> For the moment, the crunch on the amount of liquidities due to the increase in rates is not affecting data science technologies. Screening and coaching will probably still be very necessary.

> Behavioral science should be explored to see if it is truly the “silver bullet” to make interesting investments (especially in the seed stage).