Data + VC

A survey of current VC processes and recommendations for how to integrate data-driven thinking.

Justin Gage

NYU | LEONARD N. STERN SCHOOL OF BUSINESS

Abstract

The world is becoming increasingly data-driven, with traditional industries like education, healthcare, and insurance experiencing new product cycles and being challenged by startups that focus on utilizing data to improve outcomes. As citizens discuss how automation and data will render many jobs obsolete in the coming years, many argue that relationship based roles are here to stay. Investment banking is one example; Venture Capital (VC) is another.

One of the industries that believes it's safe from the incoming computer revolution is VC, a group of firms that invest institutional money into startup companies with eyes on outsized returns. Perhaps a consequence of this belief is that the industry is barely data-driven: data is rarely used in decision making and hasn't significantly transformed workflows or processes. Surprisingly, the firms that invest in the data driven future have managed to keep themselves largely out of it.

This proposal outlines how data can be used to analyze the effectiveness and success of different venture firms, with an eye on understanding what drives success and how to replicate it. As data isn't often publicly available, this project will require significant creativity to complete.

Contents

The Current State of Industry Analytics

As an industry, Venture Capital (VC) is uncharacteristically reticent when it comes to burgeoning fields of big data and analytics. There are four (4) key areas that data analytics can be utilized in VC:

A) The Hiring Process
B) Investment Sourcing
C) Investment Analysis
D) Firm Operations

We'll run through these areas and how the industry currently integrates data.

A) The Hiring Process

It's infamously difficult to get an entry level job working at a VC, for a number of reasons:

- There are likely fewer than 100 global openings per year
- Firms don't follow a set hiring schedule: they bring staff on when they need to or have the money to
- Jobs are exciting, fast paced, and strike a strong work-life balance – there are applicants from all over the place, ranging from PhDs to former founders

Firms are run by partners, and partners often have very specific preferences for what kinds of employees they want to bring in. From a number of personal conversations over the course of my time in the industry, I've heard sentiments that conflict head on.

One successful partner told me that he "would never hire someone without operating experience" and that if I ever wanted to be a good venture capitalist, I needed to spend time at a company or startup first to learn the ropes. *You can't invest in companies without having been in one for a while.* This sentiment is generally echoed by firms in Silicon Valley, while the field on the east coast is a little more mixed.

Another partner argued that if I had an opportunity to work at a VC out of college, there's no question I should take it: "VC seats don't open up often." *You get better at investing by investing and finding patterns*.

These positions conflict head on, but they share one common characteristic: they aren't based on data. They're intuitions and heuristics, and they may be right; but they aren't coming from a place of data-driven conviction. To my knowledge, nobody has collected data and analyzed how firms *are actually comprised* – do most employees of VC firms have experience working at a technology company? What percentage of partners have founded companies in the past? There has yet to be a comprehensive report detailing the experiential makeup of the VC industry.

B) Investment Sourcing

The first part of the Venture "funnel" is investment sourcing: finding startups who are raising capital and present interesting investment opportunities to the firm. Sourcing investments happens in a number of ways, including:

- Introductions from a firm's network
- Outbound communication based on an investment thesis
- References from other VC firms and accelerators

This process has been significantly improved by technology over the past few years. Data aggregation platforms like CB Insights and Pitchbook collect data on startups through algorithmic web crawlers[1] and personal phone calls to firms[2] and sell it to anyone who's interested. These platforms are prohibitively expensive for smaller venture firms, going for more than $15K per user in most cases[3]. CB Insights and Pitchbook seek to differentiate in a number of ways, including content like newsletters and blogs. In recent months a new, lower cost option has surfaced at well: CrunchBase sells a Pro subscription for $29 per month.

VCs use two general strategies to locate investment opportunities: inbound and outbound. Inbound involves emails and introductions from the VC's network – former investments, friends, colleagues, and other firms. Venture firms also engage in outbound deal flow, where they actively seek startups that they think are interesting and reach out to them as potential investments. It's in the latter category that these data aggregation platforms have been able to add value, helping VC firms to find the right companies and contact them. The industry realizes this, too. At a conference run by CB

[1] https://www.cbinsights.com/about

[2] Personal employee conversation

[3] Company websites and conversations with sales representatives

Insights last year, Sequoia Capital Partner Pat Grady argued that CB Insights "is positioned to be a leader in helping companies find investments."[4]

Helpful as the platforms have been though, they have significant shortcomings:

- Most data is aggregated from web crawlers, which means that if a company isn't on the web (or in the news), it most likely won't be on the platform
- Data on countries not in Europe or North America is totally unreliable, even for stalwart startup hubs like Israel

Pitchbook seeks to address these issues by maintaining a team of analysts that calls and verifies data with actual VC firms[5]. CB Insights has also been adding new features to the platform, like a "Mosaic" score that measures how popular a startup is[6], and a new search engine for startup and large company patents[7].

C) Investment Analysis

Deal analysis involves understanding whether an investment opportunity is a good one for a VC firm. VCs ask startups and themselves questions like:

- Is the market for this product large enough?
- Is the team uniquely capable of executing on the vision?
- Does the company have a competitive edge?

There are 3 distinct ways to integrate data into this process.

1. **Predicting Success**

The most obvious application for data in this space is creating a predictive investment model: figuring out which variables drive company performance, creating a model for the relationship, and applying that model to predict which companies are worth investing in.

This approach is largely not utilized for a number of reasons, chief one being that the variables in question are *very* difficult to quantify. If you were to include a "team" input into the model, how would you quantitatively evaluate how strong a startup's team is? Years of experience? Companies previously founded? These kinds of questions abound

[4] https://www.cbinsights.com/blog/sequoia-capital-venture-capital-disruption/

[5] Personal employee and sales representative conversations

[6] https://www.cbinsights.com/company-mosaic

[7] https://www.cbinsights.com/patent-search-engine-analytics

every area of the investment protocol, which is why the process is viewed as so subjective. Another key issue is that macroeconomic factors impact investment outcomes: there are “hot” and “cold” years for certain verticals. Success of outcomes (how large, what method) are also very difficult to quantify. We'll address these issues and seek to resolve them later.

2. **Diversification**

While hedge funds will try to diversify their investments across multiple asset classes and different industries within an asset class, VC firms rarely do the same. VCs invest largely in what comes to them (inbound) and industries that they have a thesis on. A leader in the thesis methodology is Union Square Ventures, a NYC-based firm that transparently posts all of their thoughts on the web about what they’re looking to invest in[8]. Another notable thesis-driven firm is Founders Fund, which has a “manifesto” page that details exactly what they’re looking for[9].

The process of thesis formation and the distribution of theses across different industries is intuitive and data based, but not the diversification of it. For example: a partner at USV may formulate theses about a number of industries, but that process doesn’t necessarily involve understanding how the unique financial risks of each of these industries play and interact with each other. That is to say: firms don’t attack the field by aiming to have a certain number of investments in different fields to optimally diversify risk. Partners may limit investments in a certain sector if they feel they are over-invested in it, but it’s more intuitive than data based.

3. **Deal Amount and Terms**

The amount of money invested and total round size is also a challenge to integrate into a predictive model. When should a VC “lead” a round with the first check or largest check size? When should a VC insist on certain provisions in the investment terms? These kinds of things are generally done by rules of thumb, industry standards, and instinct.

One outlier firm that takes a different approach is Correlation Ventures[10]. The firm makes very quick decisions, only follows on (does not lead investments), and uses proprietary models to decide where to invest and when. It’s difficult to get any reliable information about how exactly the firm does this, and when I spoke to a founding partner he didn’t reveal any information to me. Nevertheless, this firm is a unique exception to the rule.

---

[8] https://www.usv.com/threads

[9] http://foundersfund.com/the-future/

[10] http://correlationvc.com/

After finishing the first draft of this paper, I was pleased to be introduced to another outlier firm that takes initiative in the space: SignalFire[11]. Founded by a former Venture Partner at General Catalyst, the firm has built up a sophisticated database of hiring and investment data over the past decade[12]. Like Correlation, the details of how the fund runs are understandably kept in check. My conversation with the Founding Partner served to reinforce the analysis above.

D) Firm Operations

The final area where data can be integrated and used is in day-to-day operations of a VC firm. This typically manifests itself in two tasks: optimizing and dealing with overhead tasks (portfolio help, internal platforms) and the position of Data Scientist in Residence.

VC firms are typically small employers, but bigger funds can have many employees (more than 20). It's also a specific strategy of some firms (like Andreessen Horowitz and ffvc) to employ a lot of people to help their companies (not necessarily as investors). Maintaining many employees can require a sophisticated data system, for things like managing finances and internal research projects. I've spoken to one person who does this full time for a successful Los Angeles firm, Upfront Ventures[13]. This type of role seems to be pretty uncommon, though.

Another role that's beginning to take off is the position of Data Scientist in Residence. The DSIR completes ad hoc projects for portfolio companies that need the help of an expert, or are too young to afford to hire a full time data scientist. I've spoken to one DSIR at a well known firm, who told me that a lot of their job involves recruiting data scientists to portfolio companies.

Summary

Of the different elements of the hiring and investment processes of VCs, there is a mixed bag when it comes to data. Hiring preferences are personal and not based on any widely available data. Outbound investment activity is becoming more data-based with the rise of data aggregation platforms like CB Insights and Pitchbook, but they are limited in their scope and applicability. The decision making process is also largely intuitive and personal, without decision models or data backed methods. Firms are

---

[11] http://www.signalfire.com/

[12] https://techcrunch.com/2015/10/22/watch-out-vcs-chris-farmer-says-hes-about-to-massively-disrupt-the-industry/

[13] https://upfront.com/

beginning to hire data oriented employees to help with operations and portfolio companies.

Available Data and How to Get It

In a recent blog post, the crowdsourcing company CircleUp quoted famous Venture Capitalist Marc Andreessen in a podcast saying “machine learning wouldn’t be helpful for tech VC because there isn’t enough data.”[14] While Andreessen’s argument indeed applies to the investment model – predicting which companies to invest in – it’s less true about the other problems we’ve detailed.

The Hiring Process

Data on VC work experience should be relatively easy to acquire. Potential sources include:

- Venture firm websites and bios (for both positions and past experience)
- LinkedIn employment data
- AngelList employment data

Web scraping programs should be able to acquire this data if the correct URL is passed, and it should be relatively easy to find a comprehensive list of the top 100 or so VC firms in operation. This will create a simple data set that has all of the VC firm employees out there and the previous experience they’ve had.

There’s also potential for two more interesting datasets. The first is longitudinal: using LinkedIn and AngelList, we can view how positions have changed over time for VC firm employees. For example, if an employee was an Analyst and then promoted to Associate, that can act as two data points: an Analyst with experience set X, and an Associate with experience set X+Y.

A second potential alternative data set is the differential between public (LinkedIn and AngelList) and private (VC website) data. VC websites typically don’t show *all* of an employee’s experience, but it will be listed on public platforms. A data set can be created of all of the differences, potentially showcasing what experience VC firms value and want to show off, and those experiences that they don’t.

---

[14] http://avc.com/2017/02/machine-learning-for-investing-in-consumer-goods-startups/

The Investment Process

The investment process is indeed, as Andreessen asserted, rather opaque: it's difficult to collect and analyze anything. That being said, there are a number of ways that data can be creatively collected. There are two parts of the process to gather data on: investment and exit.

A first method is simple: using the datasets from the paid platforms CB Insights, Pitchbook, and others. These datasets include startups, money they're raised, exits, how firms co-invest, and other valuable data points – or in other words, data on both the investment and the exit processes. These sets can be used in all of the processes outlined, like sourcing investments and analyzing whether a given startup is a good investment opportunity.

Another method for collecting data is through publicly filed returns. Many VC LPs (Limited Partners) are pension funds or other vehicles that are legally obligated to disclose their investments and returns. For example, the California Public Employees' Retirement System (CALPERS) recently released figures that showed below expected returns on venture assets[15]. Venture funds are said to prefer not to take money from these types of parties that are required to disclose results, quite the indication of VC firms' desire to remain in the dark.

In terms of exits, outcomes like acquisitions and IPOs can be scraped from news sites and the like in addition to the data platforms. Other inputs that could be of interest to an investment model are macroeconomic conditions like interest rates and stock market performance, which can easily be downloaded.

[15] https://www.bloomberg.com/news/articles/2016-08-10/venture-capital-is-worst-performing-asset-for-calpers

Proposed Applications

Below I propose two practical models and projects that can help the VC industry integrate data. The first is a hiring model, and the second is an investment model. A key note to mention is that the *application* of these models is as important as the models themselves: they're proposed to work in very specific circumstances and not in others.

Hiring Analytics

HR Tech is a fast growing industry. Large corporations are trying to understand who they've been hiring, how they're been hiring them, and what has and hasn't been working. Similarly, it's important for VC firms to step back and understand who they've been hiring historically and when. It's also helpful for people in the workforce to understand what VC firms are looking for so they can get into those jobs they so covet.

After collecting the data, we can analyze it and ask questions like:

- What are the most common positions in VC?
- What are the times that firms typically hire?
- What are the positions with the most turnover?
- What experiences are most valued? Do they differ across positions?
- Which firms value which experiences over others?

The analysis can be posted on a website with accompanying graphs and explanation.

Investment Prediction Model

While the firms (Correlation and SignalFire) mentioned above have created their own proprietary models, they are just that – proprietary. My hope is that by fleshing out the core questions of model building in the public forum, more firms will feel empowered to give this thing a try. Despite the difficulty of creating a predictive investment model, it stands worth a try. Many investment models for hedge funds using alternative data started with fake money and moved their way up, so there should be no reason not to at least try it in VC.

The key reason why it *might* succeed is that I suggest applying it to a very specific stage at first.

There are massive, salient differences between VC investment stages: there's seed investing, A-B investing, growth investing, and even pre-IPO investing. These stages are dominated by different firms, different metrics, and different optimal outcomes. Growth investing, the latest venture stage, involves purchasing shares in companies after they have significant revenues. This stage is generally considered to be less risky, more stable, and most important for our purposes, more quantifiable. When deciding which companies to invest growth equity into, firms often utilize sophisticated financial models with concrete inputs. It's at this stage that our model has the highest chances of reaching initial success.

To begin thinking about how we would create and apply an investment model, we need to start off with a few key questions about how we want the model to work. Then we'll discuss the inputs we can use and outcomes to measure by.

1) Prediction or Explanation

In model building and machine learning, there's a general tradeoff between complexity and intuition – the more complex our model becomes, the less humans can understand exactly what it means. In other words, consider the following tradeoff:

A) A model that's correct 50% of the time, and clearly explains itself
B) A model that's correct 80% of the time, but reveals nothing about itself

We need to decide what we're looking for – do we want to understand the dynamics of what drives startup success? Or do we just want to create a working model that is mostly correct, even if we don't understand why? This is often dubbed the tradeoff between prediction accuracy and model interpretability.[16]

For example, we could estimate the relationship between our input variables (think: team, industry, etc.) and startup success by using a relatively simple linear regression. The model results would tell something along the lines of "*for each extra year of experience on the founding team, exit value increases by $500K*." We can interpret the model logically, and learn that experience of the founding team is indeed a driver of success.

If, on the other hand, we were to use a more complicated machine learning method like a neural net, it would be almost impossible to understand the output of the model, even though it may be more accurate.

Because this tradeoff is so steep, it makes sense to try out a few different methods and see what the results are.

[16] James, Witten, Hastie, and Tibshirani. *An Introduction to Statistical Learning With Applications in R*. New York, NY. Springer (2015). 24

2) Prediction Type

Before creating a model, we need to decide what the nature of the output will be. There are (broadly) two options:

A) *Binary Output – Yes or No*
B) *Graded Output – How Much*

A binary output is simple – after inputting data about a given startup into our model, it will spit out a decision: yes or no. The deal amount and terms are left up to the investor.

A graded output, on the other hand, suggest *how much* to invest (perhaps given a specific fund size). The decision could be binary-graded – yes or no, and if yes how much – or purely graded, with "no" decisions being judged on a certain threshold. For example, we might say that if your fund size is $20M, any investment that the model suggests under $50K is effectively a "no."

Another layer of complexity with regards to output is the terms of the investment. VC investments aren't uniform – there are "provisions" that shape the nature of the deal. For example, some VCs will negotiate rights to block a sale of the company if the offered amount is too low. Terms are important parts of the process and dispute over them can often lead to a deal falling through. In fact, the process is so complicated and obscure that a famous VC recently published a book explaining it[17].

In terms of our model, this complicates things. We could try to work terms into the model, but that would be excessively difficult. First of all, training the model would be almost impossible, as deal terms are almost never publicly available. And additionally, having the model output something along the lines of "invest in this company *if* you can get the following terms" isn't helpful.

The best compromise would be to *grade* deal terms. For example, a model output might be: "if you're able to negotiate drag-along rights, your chances of profit on this investment will increase by 10%." That way, the negotiations are left up to the investor.

3) Prediction Context

While venture investments need to be evaluated on an individual level (either binary or graded, as above), venture funds curate portfolios of investments that are *diversified*.

[17] https://www.amazon.com/Venture-Deals-Smarter-Lawyer-Capitalist/dp/1118443616

That essentially means that any given investment decision takes the existing portfolio into account – for example, a fund may not want to have more than a few companies in a given industry. We surveyed and concluded that this process isn't exactly data driven above, but it still exists nonetheless. The challenge with a model that predicts investments one by one is that it misses context – it isn't able to take those same factors into account that funds do when creating diversified portfolios. There are two major ways to address this issue.

The first is by including the existing portfolio makeup of a fund as an input into the model. The output would effectively mean: "*given a portfolio makeup of x,y,z, this is a good investment.*" This input would be difficult to measure in one metric, so in reality it would need to be a series of different inputs, like: industries invested in, stages invested in, etc.

The other way to address this issue would be to steer away from predicting whether *individual* investments are good or not in of themselves, and predict entire portfolio makeups. The output would effectively mean: "*this is a good investment, given that you also have other types of companies x,y,z in the portfolio.*" This is basically a backed-out version of the first solution and isn't very helpful. Investment opportunities present themselves one by one, not all together at once.

This is not an easy problem to solve, and speaks back to the tradeoff discussed in section 1 – between model interpretability and model accuracy. For the model to be useful going forward, it's going to need to predict the utility of investments without information about a portfolio. When backdating and analyzing past investments though, we can take the portfolio at a given time into account.

4) Measuring Success

To train our model, we'll need outcomes – end games for startups that are defined as good and bad, perhaps on a scale. An exit of $200M might be a good outcome, while an exit of $4M (or shutting down, of course) might be a poor outcome. But defining the optimal outcome isn't exactly simple.

The most obvious place to start is exit value: we could use the amount of money that a startup exits for as indicative of its success. That wouldn't work though, because the benefit *to the firm* depends on how much of the company the firm owned. But using the exit value multiplied by firm ownership *also* doesn't work, because the firm doesn't automatically take home the amount of the value that they owned – there are complicated and layered provisions (agreed upon at the time of investment) that determine how capital is payed out. Long story short, it's possible to own 10% of a company but receive less than that in certain exit scenarios.

The next logical step is to use money accrued to the firm – success can be measured by how much money the firm takes home at the end of the day. But this is also flawed, because it doesn't take firm size into account. As I wrote here and here, cash outflows to firms are defined in terms of relative success. Consider two identical scenarios for two different funds – a $200M fund and a $20M fund. Both funds own 10% of a company that exists for $100M, generally considered a relative success. For the $20M fund, they get $10M which is a great outcome (assuming they didn't invest too much) – that returns half the fund already. But for the $200M fund, that's only 5% of the fund returned. Considering that both of these funds are expected to return at least 3x the original money (net of fees), you can see that the cash flow to the firm is subjective as well.

From this analysis, it emerges that the only way to measure success has to (a) use the actual cash flow accrued to the fund, and (b) take fund size into account. Some metrics we can use would be percent of fund returned, multiple on money invested, or some function or combination of the two. In a recent comprehensive paper surveying the VC landscape, VCs themselves report that this is a common way that they assess and measure investments[18].

The final problem with measuring success is with realized vs. anticipated returns – there are many *very* successful companies that haven't returned capital yet. For example: Uber, with a valuation around $70B, is generally considered a smash hit for venture investors; yet to date it hasn't returned a dime to anyone. Our model needs to take these companies into account when determining success, perhaps with a discount applied if the company is still private. We can say that if a company's private valuation is over a certain threshold such that it can be considered a "success" – then we value that company at 85% of the valuation when it comes to our model. This threshold isn't simple to set – it can't be anything above the valuation invested at, because that doesn't ensure success. It would need to be some number that past that valuation, companies generally have a beneficial exit outcome.

A Survey of Potential Inputs

Figuring out what the inputs (features) in our model should be are beyond the scope of this paper. However, it might be helpful to list a few so that a general idea is clear. Crisply defining and quantifying these inputs is what makes the model so challenging to build.

Team

- Number of co-founders

Gompers, Paul A., William Gornall, Steven N. Kaplan, and Ilya A. Strebulaev. "How Do Venture Capitalists Make Decisions?" NBER Working Paper Series, No. 22587, September 2016.

- Age
- Work experience
- Personality traits
- Past track record
- Referrals and recommendations
- Chemistry of founding team
- Clarity of vision and passion
- Quality of other employees (non-founding) across above metrics

Market

- Current size
- Projected growth rate
- Addressable sectors
- Relevant macroeconomic factors
- Strength of competition and competitive advantage
- Existing VC investment in the space
- Acquisition landscape in the space

Product (if developed)

- Traditional and relevant KPIs (Key Performance Indicators)
- Time taken to develop
- Complexity of code base
- Reviews on the web

Clearly, the difficulty of quantifying things like “clarity of vision and passion” can be a significant deterrent to starting this project.

Conclusion

The VC industry has an opportunity to increasingly integrate data. Current processes like hiring and investing use data sparingly, but platforms are coming of age and a new wave of young VC employees will push for data to be integrated into day-to-day processes. New firms like Correlation and SignalFire will also push other VCs to ask themselves how data can be used to improve.

In this paper, I suggested two potential projects that could help VCs be better investors. The first is hiring analytics: to follow in the footsteps of big corporates and understand who generally gets hired, why, and how it worked out. The second and more ambitious project is to begin creating a predictive investment model for VC. While there are a few firms that have begun doing this, it's generally kept secret and hasn't been fleshed out well in a public space.

I'm young and relatively inexperienced, so a lot of the content in this paper may be inaccurate or lacking the proper context. That being said, the conclusions in aggregate are reasonable and were agreed to by most of the VCs that I presented these ideas to. I'm looking forward to improving and collaborating on this with anyone who's interested. My personal email is gagejustins@gmail.com.

Initial Model

To get the process started, I downloaded data on recent exits and fundraising from the Crunchbase Pro platform. The data is far from optimal, as the platform only allows you to download the 1000 most recent results, and many outcomes were missing (acquisition size, for example). Unsurprisingly, the results of initial modeling weren't promising.

After combining outcomes (IPOs and M&A) and using them as the target variable, it appears that only one variables is associated with exit values, Crunchbase Rank.

```
Call:
lm(formula = outcome ~ Crunchbase.Rank + Total.Funding.Amount +
    Number.of.Founders + Number.of.Funding.Rounds, data = comb2)

Residuals:
       Min         1Q     Median         3Q        Max
-3.172e+09 -1.192e+09 -5.424e+08  3.178e+08  2.314e+10

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              1.271e+09  9.225e+08   1.378   0.1712
Crunchbase.Rank         -2.299e+05  1.033e+05  -2.226   0.0281 *
Total.Funding.Amount     3.038e-01  4.496e-01   0.676   0.5007
Number.of.Founders       4.092e+08  2.545e+08   1.608   0.1109
Number.of.Funding.Rounds -3.697e+07  1.115e+08  -0.332   0.7408
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.067e+09 on 106 degrees of freedom
  (889 observations deleted due to missingness)
Multiple R-squared:  0.07519,   Adjusted R-squared:  0.04029
F-statistic: 2.154 on 4 and 106 DF,  p-value: 0.07916
```

The number of founders, number of funding rounds, and total funding amount have no significant association with exit values. Crunchbase Rank has a slight correlation, but being that it's partially based on exit values that's largely meaningless. The regression as a whole isn't statistically significant, with a p-value of .07 and an adjusted R^2 of 4%.

Moving forward, this model is very simplistic so there's little reason to be discouraged. With more data than just fundraising and a more robust survey of exit outcomes, it's indeed possible that there are correlations to be revealed. Other Machine Learning models like a Neural Net may also be more effective in easing out a relationship.

NYU | LEONARD N. STERN SCHOOL OF BUSINESS