VC's Framework for Evaluating an AI Startup's Tech Stack.

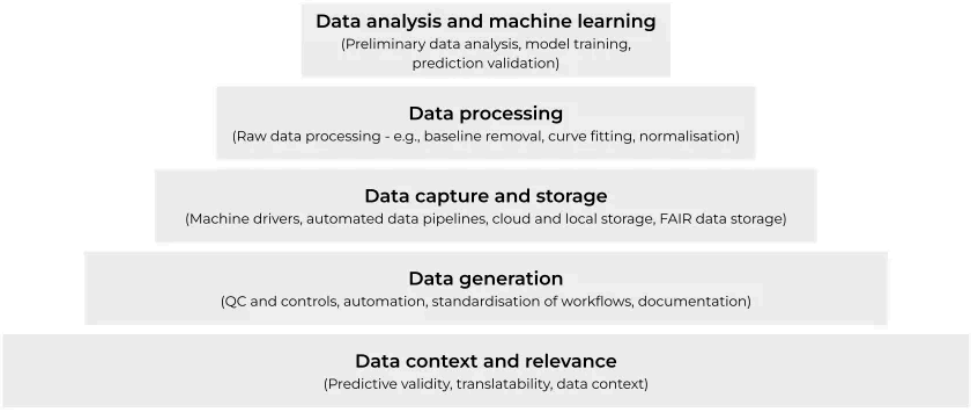
From fraud detection to agricultural crop monitoring, a new wave of tech startups has emerged, all armed with the conviction that their use of AI will address the challenges presented by the modern world.

However, as the AI landscape matures, a growing concern comes to light: The heart of many AI companies, their models, are rapidly becoming commodities. A noticeable lack of substantial differentiation among these models is beginning to raise questions about the sustainability of their competitive advantage.

Instead, while AI models continue to be pivotal components of these companies, a paradigm shift is underway. *The true value proposition of AI companies now lies not just within the models, but also predominantly in the underpinning datasets.* It is the quality, breadth, and depth of these datasets that enable models to outshine their competitors.

Many founders want to understand the framework that VCs use to evaluate the AI startup tech stack. I am sharing a framework that a General Partner shared with me.

From inconsistent datasets to noisy inputs, what could go wrong?



Written By The Venture Crew

Before jumping into the frameworks, let's first assess the basic factors that come into play when assessing data quality. And, crucially, what could go wrong if the data's not up to scratch?

Relevance

First, let's consider the datasets' relevance. Data must intricately align with the problem that an AI model is trying to solve. For instance, an AI model developed to predict housing prices necessitates data encompassing economic indicators, interest rates, real income, and demographic shifts.

Similarly, in the context of drug discovery, experimental data must exhibit the highest possible predictiveness for the effects on patients, requiring expert thought about the most relevant assays, cell lines, model organisms, and more.

Accuracy

Second, the data must be accurate. Even a small amount of inaccurate data can have a significant impact on the performance of an AI model. This is especially poignant in medical diagnoses, where a small error in the data could lead to a misdiagnosis and potentially affect lives.

Coverage

Third, coverage of data is also essential. If the data is missing important information, then the AI model will not be able to learn as effectively. For example, if an AI model is being used to translate a particular language, the data must include a variety of different dialects.

For language models, this is referred to as a "low resource" versus "high resource" language dataset. This also requires having a complete understanding of the confounding factors that affect the outcome, which typically requires the collection of metadata.

Bias

Finally, data bias also warrants rigorous consideration. Data should be captured in an unbiased way to avoid human prejudice or bias in the model. For instance, image recognition data should minimize stereotypes. In drug discovery, datasets should encompass both successful and unsuccessful molecules to avoid skewed outcomes. In both cases, the data would be considered biased and likely lose its ability to make novel predictions.

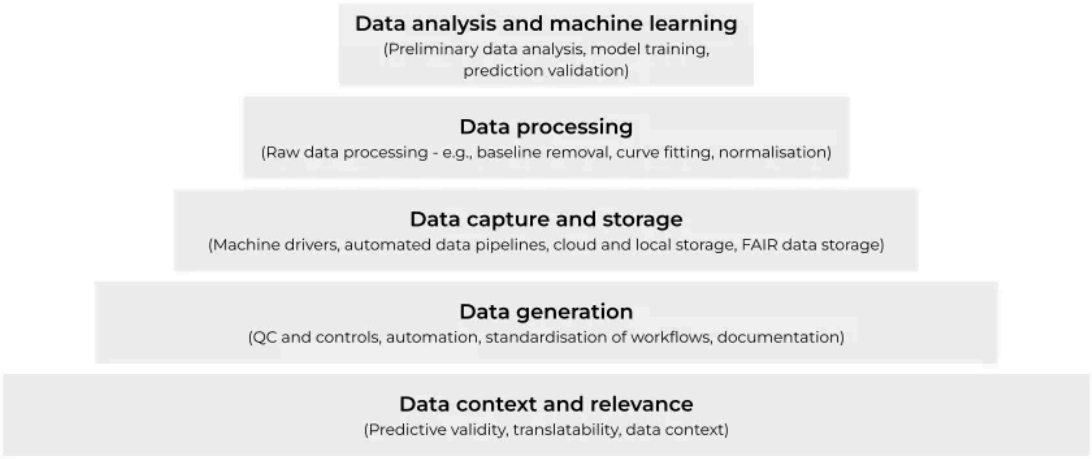
The repercussions of subpar data shouldn't be underestimated. At best, they result in a model that underperforms, and at worst, they render the model entirely ineffective. This can lead to financial losses, missed opportunities, and even physical harm.

Similarly, if the data is biased, the models will produce biased results, which can foster discrimination and unjust practices. This has been a particular concern with large language models, which have come under recent scrutiny for perpetuating stereotypes.

Compromised data quality also has the potential to erode effective decision-making, which can ultimately result in poor business performance.

Framework 1: Tech stack pyramid for data generation

To avoid investment in ineffectual AI startups, there is a need to first evaluate the processes behind the data. Picturing a company's tech stack as a pyramid is a good place to start, where the foundational tiers tend to have the biggest impact on the predictive outcome. Without this solid base, even the best data analysis and machine learning models face significant constraints.



Credit: Google Images

Here are some basic questions that a VC might initially ask to figure out if a startup's data generation process can create usable results for AI:

- Is data capture automated to enable scale-up?
- Is the data stored in secure cloud environments with automated backups?
- How is access to infrastructure and relevant compute resources managed and guaranteed?
- Are data processing pipelines fully automated, with rigorous data quality controls implemented to limit pollution from contaminated data points?

- Is the data readily accessible across the company to empower ML model-building and data-driven decisions?
- How is data governance implemented?
- Is there a data management strategy in place?
- Are data and ML model versions tracked and accessible, ensuring ML models are always working on the latest data version?

Receiving robust answers to these questions can help determine a company's grasp of the underpinning principles of its data pipelines. This understanding, in turn, will help gauge the quality of the model's output.

Framework 2: The five V's of data quality

Once a company's tech stack has been deemed suitable for AI, there is also a need to carefully consider the quality of the resulting data being used to train its models. A common framework used to capture the classification of data quality is the five V's of data quality. They represent five key dimensions of data quality that VCs should consider when evaluating AI startups:

- Veracity: The data must be accurate and truthful.
- Variety: The data must be diverse and representative of the real world.
- Volume: The data must be large enough to train the AI model effectively.
- Velocity: The data must be updated frequently to reflect changes in the world.
- Value: The data must be useful for the AI model to learn from.

Here are some introductory questions to help evaluate a company's data for the five V's:

- Does the startup have a good hypothesis about which data they need to create to build a differentiated capability or useful model?
- What data do they collect?
- Do they also collect any relevant metadata?
- How do they ensure the correctness and consistency of the data they collect?
- How does the startup plan to deal with data bias?
- Do they collect multiple examples for the same question or experiment?
- How useful is this data for the product they are building?
- What's the rationale behind collecting this data?

- Do they have evidence that their predictions improve by collecting and using this data? If yes, how does the data amount correlate with prediction improvement
- How easy is it for a competitor to collect the same data? How long would it take and how much would it cost for them to do so?

- Specifically for a biotech, how well does the proxy they are predicting correlate with a clinically relevant endpoint? Is there evidence for this?
- What is the startup's plan for ensuring the quality of its data over time?
- How does the startup plan to protect its data from unauthorized access?
- How does the startup plan to comply with data privacy regulations?

By carefully considering the five V's of data quality, VCs can make sure they are investing in AI startups that have the data they need to succeed. If the startup can answer the above questions convincingly and their data scores highly in the five dimensions, it is a good sign that they are serious about data quality and are properly equipped to apply their AI models.

Finally, VCs should assess the startup's commitment to data security. This includes things like their data governance policies, their data quality assurance procedures, and their data breach response plans.

Interrogate the hype to find the winners

Amid the resounding buzz surrounding AI in recent months, the allure of substantial investments has attracted startup founders willing to exaggerate their infrastructure and inflate capabilities in the search for capital.

Successful VCs are asking the right questions to interrogate these companies thoroughly and filtering out the potential winners built on a solid foundation from those with a hollow shell that are ultimately destined to fail.
