Taken from: https://a16z.com/2020/02/16/the-new-business-of-ai-and-how-its-different-from-traditional-software/

The New Business of AI (and How It's Different From Traditional Software)

by Martin Casado and Matt Bornstein

At a technical level, artificial intelligence seems to be the future of software. AI is showing remarkable progress on a range of difficult computer science problems, and the job of software developers – who now work with data as much as source code – is changing fundamentally in the process.

Many AI companies (and investors) are betting that this relationship will extend beyond just technology – that AI *businesses* will resemble traditional software companies as well. Based on our experience working with AI companies, we're not so sure.

We are huge believers in the power of AI to transform business: We've put our money behind that thesis, and we will continue to invest heavily in both applied AI companies and AI infrastructure. However, we have noticed in many cases that AI companies simply don't have the same economic construction as software businesses. At times, they can even look more like traditional services companies. In particular, many AI companies have:

1. **Lower gross margins** due to heavy cloud infrastructure usage and ongoing human support;
2. **Scaling challenges** due to the thorny problem of edge cases;
3. **Weaker defensive moats** due to the commoditization of AI models and challenges with data network effects.

Anecdotally, we have seen a surprisingly consistent pattern in the financial data of AI companies, with gross margins often in the 50-60% range – well below the 60-80%+ benchmark for comparable SaaS businesses. Early-stage private capital can hide these inefficiencies in the short term, especially as some investors push for growth over profitability. It's not clear, though, that any amount of long-term product or go-to-market (GTM) optimization can completely solve the issue.

Just as SaaS ushered in a novel economic model compared to on-premise software, we believe *AI is creating an essentially new type of business*. So this post walks through some of the ways AI companies differ from traditional software companies and shares some advice on how to address those differences. Our goal is not to be prescriptive but rather help operators and others understand the economics and strategic landscape of AI so they can build enduring companies.

We have noticed in many cases that AI companies simply don't have the same economic construction as software businesses. At times, they can even look more like traditional services companies. Click To Tweet

Software + services = AI?

The beauty of software (including SaaS) is that it can be produced once and sold many times. This property creates a number of compelling business benefits, including recurring revenue streams, high (60-80%+) gross margins, and – in relatively rare cases when network effects or scale effects take hold – *superlinear* scaling. Software companies also have the potential to build strong defensive moats because they own the intellectual property (typically the code) generated by their work.

Service businesses occupy the other end of the spectrum. Each new project requires dedicated headcount and can be sold exactly once. As a result, revenue tends to be non-recurring, gross margins are lower (30-50%), and scaling is linear at best. Defensibility is more challenging – often based on brand or incumbent account control – because any IP not owned by the customer is unlikely to have broad applicability.

AI companies appear, increasingly, to combine elements of both software and services.

Most AI applications look and feel like normal software. They rely on conventional code to perform tasks like interfacing with users, managing data, or integrating with other systems. The heart of the application, though, is a set of trained data models. These models interpret images, transcribe speech, generate natural language, and perform other complex tasks. Maintaining them can feel, at times, more like a services business – requiring significant, customer-specific work and input costs beyond typical support and success functions.

This dynamic impacts AI businesses in a number of important ways. We explore several – gross margins, scaling, and defensibility – in the following sections.

AI companies appear, increasingly, to combine elements of both software and services with gross margins, scaling, and defensibility that may represent a new class of business entirely. Click To Tweet

Gross Margins, Part 1: Cloud infrastructure is a substantial – and sometimes hidden – cost for AI companies

In the old days of on-premise software, delivering a product meant stamping out and shipping physical media – the cost of running the software, whether on servers or desktops, was borne by the buyer. Today, with the dominance of SaaS, that cost has been pushed back to the vendor. Most software companies pay big AWS or Azure bills every month – the more demanding the software, the higher the bill.

AI, it turns out, is pretty demanding:

- **Training** a single AI model can cost hundreds of thousands of dollars (or more) in *compute* resources. While it's tempting to treat this as a one-time cost, retraining is increasingly recognized as an ongoing cost, since the data that feeds AI models tends to change over time (a phenomenon known as "data drift").
- **Model inference** (the process of generating predictions in production) is also more computationally complex than operating traditional software. Executing a long series of matrix multiplications just requires more math than, for example, reading from a database.
- AI applications are more likely than traditional software to operate on **rich media** like images, audio, or video. These types of data consume higher than usual *storage* resources, are expensive to process, and often suffer from region of interest issues – an application may need to process a large file to find a small, relevant snippet.
- We've had AI companies tell us that **cloud operations** can be more complex and costly than traditional approaches, particularly because there aren't good tools to scale AI models globally. As a result, some AI companies have to routinely transfer trained models across cloud regions – racking up big ingress and egress costs – to improve reliability, latency, and compliance.

Taken together, these forces contribute to the 25% or more of revenue that AI companies often spend on cloud resources. In extreme cases, startups tackling particularly complex tasks have actually found manual data processing cheaper than executing a trained model.

Help is coming in the form of specialized AI processors that can execute computations more efficiently and optimization techniques, such as model compression and cross-compilation, that reduce the number of computations needed.

But it's not clear what the shape of the efficiency curve will look like. In many problem domains, exponentially more processing and data are needed to get incrementally more accuracy. This means – as we've noted before – that model complexity is growing at an incredible rate, and it's unlikely processors will be able to keep up. Moore's Law is not enough. (For example, the compute resources required to train state-of-the-art AI models has grown over 300,000x since 2012, while the transistor count of NVIDIA GPUs has grown only ~4x!) Distributed computing is a compelling solution to this problem, but it primarily addresses speed – not cost.

Gross Margins, Part 2: Many AI applications rely on "humans in the loop" to function at a high level of accuracy

Human-in-the-loop systems take two forms, both of which contribute to lower gross margins for many AI startups.

First: training most of today's state-of-the-art AI models involves the manual cleaning and labeling of large datasets. This process is laborious, expensive, and among the biggest barriers to more widespread adoption of AI. Plus, as we discussed above, training doesn't end once a model is deployed. To maintain accuracy, new training data needs to be continually captured, labeled, and fed back into the system. Although techniques like drift detection and active learning can reduce the burden, anecdotal data shows that many companies spend up to 10-15% of revenue on this process – usually not counting core engineering resources – and suggests ongoing development work exceeds typical bug fixes and feature additions.

Second: for many tasks, especially those requiring greater cognitive reasoning, humans are often plugged into AI systems in real time. Social media companies, for example, employ thousands of human reviewers to augment AI-based moderation systems. Many autonomous vehicle systems include remote human operators, and most AI-based medical devices interface with physicians as joint decision makers. More and more startups are adopting this approach as the capabilities of modern AI systems are becoming better understood. A number of AI companies that planned to sell pure software products are increasingly bringing a services capability in-house and booking the associated costs.

The need for human intervention will likely decline as the performance of AI models improves. It's unlikely, though, that humans will be cut out of the loop entirely. Many problems – like self-driving cars – are too complex to be fully automated with current-generation AI techniques. Issues of safety, fairness, and trust also demand meaningful human oversight – a fact likely to be enshrined in AI regulations currently under development in the US, EU, and elsewhere.

The need for human intervention will likely decline as the performance of AI models improves. It's unlikely, though, that humans will be cut out of the loop entirely. Click To Tweet

Even if we do, eventually, achieve full automation for certain tasks, it's not clear how much margins will improve as a result. The basic function of an AI application is to process a stream of input data and generate relevant predictions. The cost of operating the system, therefore, is a function of the amount of data being processed. Some data points are handled by humans (relatively expensive), while others are processed automatically by AI models (hopefully less expensive). But every input needs to be handled, one way or the other.

For this reason, the two categories of costs we've discussed so far – cloud computing and human support – are actually linked. Reducing one tends to drive an increase in the other. Both pieces of the equation can be optimized, but neither one is likely to reach the near-zero cost levels associated with SaaS businesses.

Scaling AI systems can be rockier than expected, because AI lives in the long tail

For AI companies, knowing when you've found product-market fit is just a little bit harder than with traditional software. It's deceptively easy to think you've gotten there – especially after closing 5-10 great customers – only to see the backlog for your ML team start to balloon and customer deployment schedules start to stretch out ominously, drawing resources away from new sales.

The culprit, in many situations, is edge cases. Many AI apps have open-ended interfaces and operate on noisy, unstructured data (like images or natural language). Users often lack intuition around the product or, worse, assume it has human/superhuman capabilities. This means edge cases are everywhere: as much as 40-50% of intended functionality for AI products we've looked at can reside in the long tail of user intent.

Put another way, users can – and will – enter just about anything into an AI app.

Users can – and will – enter just about anything into an AI app. Click To Tweet

Handling this huge state space tends to be an ongoing chore. Since the range of possible input values is so large, each new customer deployment is likely to generate data that has never been seen before. Even customers that appear similar – two auto manufacturers doing defect detection, for example – may require substantially different training data, due to something as simple as the placement of video cameras on their assembly lines.

One founder calls this phenomenon the "time cost" of AI products. Her company runs a dedicated period of data collection and model fine-tuning at the start of each new customer engagement. This gives them visibility into the distribution of the customer's data and eliminates some edge cases prior to deployment. But it also entails a cost: the company's team and financial resources are tied up until model accuracy reaches an acceptable level. The duration of the training period is also generally unknown, since there are typically few options to generate training data faster… no matter how hard the team works.

AI startups often end up devoting more time and resources to deploying their products than they expected. Identifying these needs in advance can be difficult since traditional prototyping tools – like mockups, prototypes, or beta tests – tend to cover only the most common paths, not the edge cases. Like traditional software, the process is especially time-consuming with the earliest customer cohorts, but unlike traditional software, it doesn't necessarily disappear over time.

The playbook for defending AI businesses is still being written

Great software companies are built around strong defensive moats. Some of the best moats are strong forces like network effects, high switching costs, and economies of scale.

All of these factors are possible for AI companies, too. The foundation for defensibility is usually formed, though – especially in the enterprise – by a technically superior product. Being the first to implement a complex piece of software can yield major brand advantages and periods of near-exclusivity.

In the AI world, technical differentiation is harder to achieve. New model architectures are being developed mostly in open, academic settings. Reference implementations (pre-trained models) are available from open-source libraries, and model parameters can be optimized automatically. Data is the core of an AI system, but it's often owned by customers, in the public domain, or over time becomes a commodity. It also has diminishing value as markets mature and shows relatively weak network effects. In some cases, we've even seen *diseconomies* of scale associated with the data feeding AI businesses. As models become more mature – as argued in "The Empty Promise of Data Moats" – each new edge case becomes more and more costly to address, while delivering value to fewer and fewer relevant customers.

This does not necessarily mean AI products are less defensible than their pure software counterparts. But the moats for AI companies appear to be shallower than many expected. AI may largely be a pass-through, from a defensibility standpoint, to the underlying product and data.

This does not necessarily mean AI products are less defensible than their pure software counterparts. But the moats for AI companies appear to be shallower than many expected. Click To Tweet

Building, scaling, and defending great AI companies – practical advice for founders

We believe the key to long-term success for AI companies is to own the challenges and combine the best of both services and software. In that vein, here are a number of steps founders can take to thrive with new or existing AI applications.

Eliminate *model complexity* as much as possible. We've seen a massive difference in COGS between startups that train a unique model per customer versus those that are able to share a single model (or set of models) among all customers. The "single model" strategy is easier to maintain, faster to roll out to new customers, and supports a simpler, more efficient engineering org. It also tends to reduce data pipeline sprawl and duplicative training runs, which can meaningfully improve cloud infrastructure costs. While there is no silver bullet to reaching this ideal state, one key is to understand as much as possible about your customers – and their data – *before* agreeing to a deal. Sometimes it's obvious that a new customer will cause a major fork in your ML engineering efforts. Most of the time, the changes are more subtle, involving only a few unique models or some fine-tuning. Making these judgment calls – trading off long-term economic health versus near-term growth – is one of the most important jobs facing AI founders.

Choose problem domains carefully – and often narrowly – to reduce *data complexity*. Automating human labor is a fundamentally hard thing to do. Many companies are finding that the minimum viable task for AI models is narrower than they expected. Rather than offering general text suggestions, for instance, some teams have found success offering short suggestions in email or job postings. Companies working in the CRM space have found highly valuable niches for AI based just around updating records. There is a large class of problems, like these, that are hard for humans to perform but relatively easy for AI. They tend to involve high-scale, low-complexity tasks, such as moderation, data entry/coding, transcription, etc. Focusing on these areas can minimize the challenge of persistent edge cases – in other words, they can simplify the data feeding the AI development process.

We believe the key to long-term success for AI companies is to own the challenges and combine the best of both services and software. Click To Tweet

Plan for high variable costs. As a founder, you should have a reliable, intuitive mental framework for your business model. The costs discussed in this post are likely to get better – reduced by some constant – but it would be a mistake to assume they will disappear completely (or to force that unnaturally). Instead, we suggest building a business model and GTM strategy with lower gross margins in mind. Some good advice from founders: Understand deeply the distribution of data feeding your models. Treat model maintenance and human failover as first-order problems. Track down and measure your real variable costs – don't let them hide in R&D. Make conservative unit economic assumptions in your financial models, especially during a fundraise. Don't wait for scale, or outside tech advances, to solve the problem.

Embrace services. There are huge opportunities to meet the market where it stands. That may mean offering a full-stack translation service rather than translation software or running a taxi service rather than selling self-driving cars. Building hybrid businesses is harder than pure software, but this approach can provide deep insight into customer needs and yield fast-growing, market-defining companies. Services can also be a great tool to kickstart a company's go-to-market engine – see this post for more on this – especially when selling complex and/or brand new technology. The key is pursue one strategy in a committed way, rather than supporting both software and services customers.

Plan for change in the tech stack. Modern AI is still in its infancy. The tools that help practitioners do their jobs in an efficient and standardized way are just now being built. Over the next several years, we expect to see widespread availability of tools to automate model training, make inference more efficient, standardize developer workflows, and monitor and secure AI models in production. Cloud computing, in general, is also gaining more attention as a cost issue to be addressed by software companies. Tightly coupling an application to the current way of doing things may lead to an architectural disadvantage in the future.

Build defensibility the old-fashioned way. While it's not clear whether an AI model itself – or the underlying data – will provide a long-term moat, good products and proprietary data almost always builds good businesses. AI gives founders a new angle on old problems. AI techniques, for example, have delivered novel value in the relatively sleepy malware detection market by simply showing better performance. The opportunity to build sticky products and enduring businesses on top of initial, unique product capabilities is evergreen. Interestingly, we've also seen several AI companies cement their market position through an effective cloud strategy, similar to the most recent generation of open-source companies.

* * *

To summarize: most AI systems today aren't *quite* software, in the traditional sense. And AI businesses, as a result, don't look exactly like software businesses. They involve ongoing human support and material variable costs. They often don't scale quite as easily as we'd like. And strong defensibility – critical to the "build once / sell many times" software model – doesn't seem to come for free.

These traits make AI feel, to an extent, like a services business. Put another way: you can replace the services firm, but you can't (completely) replace the services.

Believe it or not, this may be good news. Things like variable costs, scaling dynamics, and defensive moats are ultimately determined by markets – not individual companies. The fact that we're seeing unfamiliar patterns in the data suggests AI companies are truly something new – pushing into new markets and building massive opportunities. There are already a number of great AI companies who have successfully navigated the idea maze and built products with consistently strong performance.

Things like variable costs, scaling dynamics, and defensive moats are ultimately determined by markets – not individual companies. The fact that we're seeing unfamiliar patterns in the data suggests AI companies are truly something new –… Click To Tweet

AI is still early in the transition from research topic to production technology. It's easy to forget that AlexNet, which arguably kickstarted the current wave of AI software development, was published less than eight years ago. Intelligent applications are driving the software industry forward, and we're excited to see where they go next.

—

Sources: Gross margin estimates for traditional software were based on a selection of companies listed on publiccomps.com; gross margin estimates for services companies were based on 10k filings; and gross margin estimates for AI businesses were based on several interviews with founders of AI startups.

The views expressed here are those of the individual AH Capital Management, L.L.C. ("a16z") personnel quoted and are not the views of a16z or its affiliates. Certain information contained in here has been obtained from third-party sources, including from portfolio companies of funds managed by a16z. While taken from sources believed to be reliable, a16z has not independently verified such information and makes no representations about the enduring accuracy of the information or its appropriateness for a given situation.

This content is provided for informational purposes only, and should not be relied upon as legal, business, investment, or tax advice. You should consult your own advisers as to those matters. References to any securities or digital assets are for illustrative purposes only, and do not constitute an investment recommendation or offer to provide investment advisory services. Furthermore, this content is not directed at nor intended for use by any investors or prospective investors, and may not under any circumstances be relied upon when making a decision to invest in any fund managed by a16z. (An offering to invest in an a16z fund will be made only by the private placement memorandum, subscription agreement, and other relevant documentation of any such fund and should be read in their entirety.) Any investments or portfolio companies mentioned, referred to, or described are not representative of all investments in vehicles managed by a16z, and there can be no assurance that the investments will be profitable or that *other investments made in the future will have similar characteristics or results. A list of investments made by funds managed by Andreessen Horowitz (excluding investments for which the issuer has not provided permission for a16z to disclose publicly as well as unannounced investments in publicly traded digital assets) is available at https://a16z.com/investments/.*

Charts and graphs provided within are for informational purposes solely and should not be relied upon when making any investment decision. Past performance is not indicative of future results. The content speaks only as of the date indicated. Any projections, estimates, forecasts, targets, prospects, and/or opinions expressed in these materials are subject to change without notice and may differ or be contrary to opinions expressed by others. Please see https://a16z.com/disclosures for additional important information.