

Everything, Everywhere, All AI

Bessemer's investment
strategies for the
AI revolution

A book by
Bessemer Venture Partners

Copyright © 2024

Atlas reads

Inside Bessemer's operating model 3
Featuring Jeremy Levine, Brian Feinstein, Kent Bennett

Investment roadmaps and trends

Big Tech's battle over AI foundation models 11
Janelle Teng, Sameer Dholakia

AI Infrastructure 15
Janelle Teng, Grace Ma, Bhavik Nagda,
Mary D'Onofrio, Elliott Robinson, David Cowan

AI makes all of us 10x developers 22
Lindsey Li, Bhavik Nagda

The data shift right market 26
Lindsey Li, Yael Schiff, Amit Karp

Multimodal innovation transforms human relationships with software 31
Aia Sarycheva, Mike Droesch

Vertical AI dwarfs legacy vertical SaaS with new applications and business models 35
Caty Rea, Brian Feinstein

AI brings Consumer Cloud back 40
Maha Malik, Talia Goldberg, Kent Bennett

Atlas reads

The autonomous robotics future is around the corner 44
Alex Ferrara, Aditya Nidmarti, Bhavik Nagda

Unlocking machine learning for drug discovery 52
Andrew Hedin

Lessons for AI leaders

Six imperatives for AI-first companies 63
Morgan Cheatham and Steve Kraus

**Seven product strategies to prevent churn
for B2B AI app leaders** 74
Janelle Teng and Sameer Dholakia

**How Intercom navigated the
AI paradigm shift** 83

**AI escape velocity: A conversation between
Ray Kurzweil and Talia Goldberg** 88

**Crossing The Rubicon:
Our \$1 billion commitment to AI** 100

Inside Bessemer's operating model

Partners Jeremy Levine, Kent Bennett, and Brian Feinstein tell Bessemer's origin story and explain how it shapes the way Bessemer makes investments today — embracing collaboration, intellectual honesty, and rethinking traditions.



Venture insights that matter

bvp.com/subscribe



From steel to AI — The paradox, principles, and approach to our unique investment strategy

With over 20 investing partners and nine offices around the globe, Bessemer's footprint extends from Silicon Valley, New York and Cambridge all the way to London, Israel, Bangalore, and beyond. So what's the secret to the operating model that runs one of venture capital's most tenured institutions? Paradox.

Bessemer is highly decentralized, driven by independent thinkers, and yet, remains a close knit partnership that holds intellectual honesty and collaboration as core values in its code of conduct. Because the firm doesn't have a single owner at the helm, partners can make decisions autonomously while still benefiting from the collective wisdom of peers and mentors, as well as insights from past generations of Bessemer investors.

"We work in a highly disaggregated, empowered manner where any partner who puts in the time, effort, and capital can make investments in what they believe will become the next great technological invention, business, or trend," explains Jeremy Levine, a partner who has spent more than two decades in Bessemer's New York City office.

"Autonomy is so important to us," adds Bessemer Partner Brian Feinstein. "Conventional wisdom says you can't scale venture capital, and that's often true with a consensus decision-making model. We like to think we've cracked the code on scaling by enabling each partner to pursue investment areas as they see fit and make the best decisions they possibly can."

Within this model, questioning the status quo isn't just allowed, it's seen as a positive signal of the firm's healthy functioning. "In the 15 years I've been at Bessemer, our operating system has changed in major ways, multiple times," says Kent Bennett, a partner who's invested in software at Bessemer since 2008. "Our most long-standing tradition is actually rethinking traditions."

Bessemer has always valued reflection and transparency — in fact, we even open-source our research on the industries and trends where we invest, as well as our investment memos from the last quarter century. But it's still rare to hear three partners go in-depth on the dynamics and values that have defined Bessemer for over a century.

Here, Jeremy Levine, Brian Feinstein, and Kent Bennett discuss the firm's history, evolution, and operating model, and the impact of those foundations on the firm's culture, talent, and investments. The partners also share what they look for in companies, why they've committed \$1 billion to AI investments, and other VC insights from a combined 55+ year tenure at Bessemer.

A brief history of Bessemer

Bessemer origins trace back to the family office of one of the lesser known co-founders of Carnegie Steel, Henry Phipps. "None of us were even born when what became Bessemer Venture Partners started, but its history has been passed down over the years and decades," says Jeremy.

"Back in the 1890s, Andrew Carnegie, Henry Phipps and Henry Frick commercialized the Bessemer process — a then newly invented way of making steel at high volumes very cost effectively. Phipps took his earnings and put them into a trust for his descendants, with a goal of creating a really long-term family office."

In 1975, the department of Phipps's family office responsible for making high-risk, high-growth investments began focusing on high-tech projects, opened its office in Silicon Valley, and officially became Bessemer Venture Partners.



"The family understood that performance will go naturally up and down over years and decades, which allowed Bessemer to invest in five, 10, and 20-year projects. That ultimately drove fantastic performance and created the stable capital base that we benefit from today," explains Jeremy.

That's not the only influence of Phipps's legacy on the firm. "Phipps was a private individual and named his own family company Bessemer in a nod to the brilliant inventor who created the process that generated all of his wealth," explains Jeremy. "The focal point was always the technology, not the individual, and that's very much in line with our ethos today."

Bessemer's unique operating model

With no living founder, Bessemer's operating model is set up to endure past any one individual. The firm's commitment to autonomy and independent thinking has driven its longevity through market cycles and decades of technological advancements.

"It's unique that there's no one partner in charge of investment decisions at Bessemer. In the 22 years I've been at the firm, I've seen just one investment get shot down," recalls Jeremy. "We don't need to ask anyone permission. Instead, we ask each other for feedback on investment decisions that we have the authority to make and the responsibility to make well."

Each partner is free to make investments based on their convictions and roadmap strategies.

In other words, months and years of deep research in emerging areas of technology — from cloud and quantum to healthcare and artificial intelligence — underpin the decisions to back certain audacious founders and startups. In weekly partnership meetings, investors present opportunities, share memos and due diligence, and often have the founders present to the firm. Then, in a roundtable discussion, others ask questions and share feedback and recommendations to the lead investor. Given the autonomy each partner has in driving investments, these meetings are not a function to get permission. The decision to invest rests to a large extent with the partner and the teaming working on the deal. However, honest feedback helps sharpen each others's thinking and roadmap development.

"We're asking each other for feedback on what we individually want to do, so we feel the responsibility and the authority to make the decision," explains Jeremy. "We want to be able to tell each other gee Brian or gee Kent or gee Jeremy, that doesn't look like a very good decision you're making."

These rigorous discussions culminate in a 1-10 vote that represents collective feedback on the investment, not necessarily an official ruling.

The impacts of this decision-making model are diverse and wide reaching. During their tenures, Jeremy, Kent, and Brian have observed how it has shaped culture, investment outcomes, talent, and relationships, both among colleagues at Bessemer and with founders and other operators.

Intellectual honesty

"One of the first things you'll notice at Bessemer is that people just say what they are thinking. There's no politics or salesmanship because there's no founder of the corner office that you have to worry about offending by giving your honest opinion," says Kent.

Jeremy recalls an onboarding meeting with a new employee where he asked if anything had been a departure from what she had expected, and she immediately brought up Bessemer's investment memos.



"She couldn't believe how honest they were. At her previous firm, [memos](#) were written to convince those who ran the investment committee. In contrast, it was very clear to her that the partners at Bessemer wrote the memos to really explain their decision, not to ask someone else to make that decision for them."

Investor tenure

Many investors — Brian, Jeremy, and Kent among them — joined the firm early in their careers, received mentorship from senior partners, and have since paid that mentorship forward to newer recruits over the course of their long tenures.

"I spent several years listening to my first mentor, Felda Hardymon, talk to companies," recalls Kent. "I sat next to Jeremy on two boards. I sat next to Bob Goodman, a senior partner, for 200 hours of meetings. There's no shortcut to learning what you learn over that amount of time, surrounded by people with that level of expertise and experience."

This apprenticeship model and the ability to have ownership over investment decisions attracts new generations of talent to Bessemer. "As a young person growing up here, it's wildly exciting and empowering to know you'll get to establish your own track record and reputation," says Jeremy.

Investment outcomes

Investors' ability to make their own decisions based on research, expertise, and past experience allows for more variability in the type of investments being made and reduces the chance that great but unconventional investments are passed over.

"Over time, each partner picks up over time a number of essential themes — we call them roadmaps — and does a bunch of homework on them. Then, they bring the insights to the team and start investing in companies," explains Kent.

"You quickly see that the goal is not to invest in what's popular," adds Brian. "It's not about chasing momentum. It's not about pursuing the consensus opportunities. It's about developing a thesis and pursuing ideas that might be off the beaten path.

"Three of our most successful vertical software investments — Shopify, Toast and Procore — were all relatively unpopular investments when we first discussed them in a partnership meeting. Our autonomous decision-making model made them possible."

Improved relationships

Investor autonomy also leads to more transparency in relationships — both among partners, who can give and get honest opinions, and with the portfolio companies that they invest in, who save time and get more hands-on support because they're working directly with a decision-maker.

"When I show interest in investing in companies, founders usually expect they'll have to jump through a series of hoops back at the firm," says Kent. "I love getting to tell the actual process: 'We meet, and then I'll go back to the team and decide whether we want to invest.'"

"You get similar benefits when you sit on the board of a company and the company is facing a tough decision. The co-investors around the table will say things like, 'I have to go ask my partners what I can do here.' It's really empowering as an investor to be able to just say, 'I'm the one with the most context, and here's what we can do.'"



Distinct styles, collective approach: Bessemer's investment philosophy

The principle of autonomy extends to partners' investment interests. "Many associate Bessemer with software, but we also have investments in life sciences, frontier tech, fintech, and beyond. We make speculative seed investments and late-stage growth buyouts of mature companies generating profit. The breadth of our investments is really extraordinary," says Brian.

What each partner weighs most in investments varies too. "The thing I fall hardest for is wild capital efficiency," says Jeremy. "At the earliest stages, that manifests as scrappiness and the willingness to do things that are the opposite of gold plated. That's been the case for many of my investments. But I believe it's important to be willing to break your own rules. One of my most successful investments, Pinterest, ended up raising \$1 billion of equity capital before its IPO."

Kent, who invests primarily on B2B software, is laser-focused on product differentiation. "The investments that succeeded had products that were clearly superior to the existing alternatives. The ones that failed had a product that was more of a question mark — either because it wasn't fully formulated or because it wasn't a slam dunk in the eyes of consumers. Over time, I've honed an ability to identify products that are radically advantaged from day one."

Brian assesses companies based on their path to market leadership and defensibility. "In vertical software, there are these virtuous cycles where first movers have a huge advantage. In terms of founders, I tend to work with leaders who have a deep empathy for the domain and customers. In vertical software, I find this translates to long-term success."

Each partner has specific stages, geographies, and roadmaps where they might have more expertise and therefore invest in more. Still, no one lays claim to any one area. The firm has a strong culture of sharing and seeking feedback, so ideas are circulated widely, and when a fertile area of investment is discovered, other partners typically get involved.

"A great example is Brian's investment in Mindbody. He discovered the company when he was an analyst and I was a partner, and we ended up investing," recalls Jeremy. "When we realized that Mindbody's payments product for fitness and wellness businesses could drive significant revenue growth, we tried to find every other company doing something similar in other categories

"Ultimately, we had 70 vertical software companies invested in by nine different partners. If one partner had called it off limits (to the others), we would have been maybe 10% as productive. However, it's our practice to help each other by sharing our ideas and research, and benefit from each other's execution against the best ideas."

Adaptation to technological shifts, especially in the AI economy

Venture is a game of spotting the technological waves before they crest and, eventually, define a generation. This strategy not only informs Bessemer's investments in new areas but also ensures its portfolio companies leverage these advancements to remain competitive and hopefully grow into market leaders. Many partners currently at Bessemer were investors during recent waves — cloud, mobile, and vertical software — witnessed first-hand how these shifts transformed the world's digital economy.



"With the benefit of hindsight, the importance and massive impact of new technologies like cloud and mobile are obvious. But when they first emerged, I'm not sure they felt all that different from VR and blockchain. All these things feel incredibly exciting in the moment, and we've seen some of them make it, while others don't," says Jeremy.

As a leader in the cloud revolution, Bessemer believes tech is now transitioning from an era of CPUs to GPUs, in the dawn of the AI-era. In 2023, Bessemer announced its \$1 billion commitment to backing AI-Native founders everywhere: "We're amidst a new and major computing revolution," the firm shared.

"Artificial intelligence is here and nearing escape velocity. Progress across numerous technological vectors has led us to this point — from new model architectures to specialized hardware with vast computing power to advanced machine learning techniques. There has never been a better time for small, ambitious teams to positively transform life as we know it."

"The arrival of advanced LLMs is relatively new, but we're very enthusiastic. We think it's much more likely than not to have a huge impact on the industries where we invest," says Jeremy.

Kent adds, "I see AI as another golden hammer in the tool belt of companies looking to create incredible software. Companies can use it to accelerate their efficiency and make the company more productive per person, and they can also use it to build the best possible product.

"I don't necessarily think that AI is more likely than other technologies to create a network effect — though some make good arguments that it could — but I do think that companies that leverage the technology will be able to build an even more jaw-dropping product, get it to market faster, and provide better service, all of which drive long-term defensibility."

Since ChatGPT's astonishing release into the world, portfolio companies such as Intercom, Zapier, and Canva have demonstrated how category leaders are adapting to the AI imperative by enabling their software platforms with LLM and AGI advancements.

Recent Native AI leaders have joined Bessemer portfolio, including Abridge, Anthropic, DeepL, Coactive, EvenUp, Fieldguide, and Perplexity, among others. Our partners continue to research and make predictions on ways subsectors of the AI economy could evolve — from foundational models and AI infrastructure to B2B and Vertical AI businesses, and beyond.



Stewards of capital, optimists of the future

Longevity sits at the center of everything that Bessemer is and does — from the long tenures of the partner to a stable capital base built from over a century of investments. Key to longevity in venture capital however, is staying open to change, proactively seeking out fresh and contrarian perspectives, and embracing the longview.

"We feel incredibly fortunate to have been given this platform by the prior generation," says Brian. "When someone hands something over to you, you're the beneficiary of it, and you're able to build a great career and do incredible things. You just feel an enormous responsibility to hand it over to the next generation.

"There is no 'one way' to be successful in venture capital. You don't need to be like anyone else. You just need to find your path, and discover the way you can do it better than everyone else. Play to your strengths and be authentic to your character."

The original interview that inspired this article first aired on the podcast [Invest Like The Best](#).

Investment roadmaps & trends



Venture insights that matter

bvp.com/subscribe

Big Tech's battle over AI foundation models

Janelle Teng, Sameer Dholakia



Venture insights that matter

bvp.com/subscribe

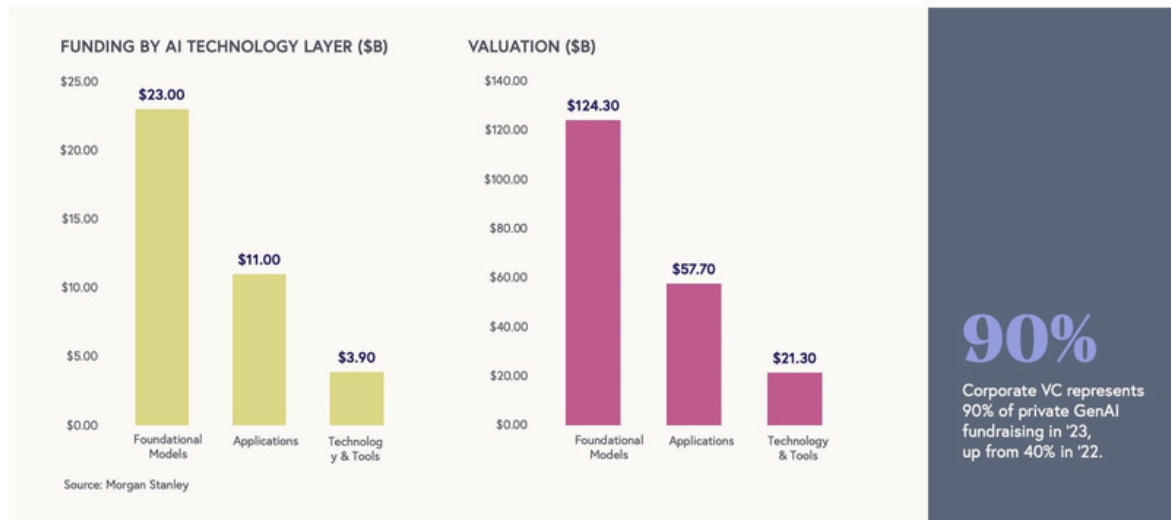


AI foundation models set the stage for Big Tech's new battle-of-the-century

When we reflect on the platform shifts of yesteryear — from Browsers and Search to Mobile and Cloud — every technological change catalyzed competition to control the foundational layer. The age of AI is no different. Foundation models are the new "oil" that will fuel downstream AI applications and tooling.

AI value creation is currently concentrated at the model layer

In 2023, foundational model companies captured accounting for over 60% of the total VC dollars.



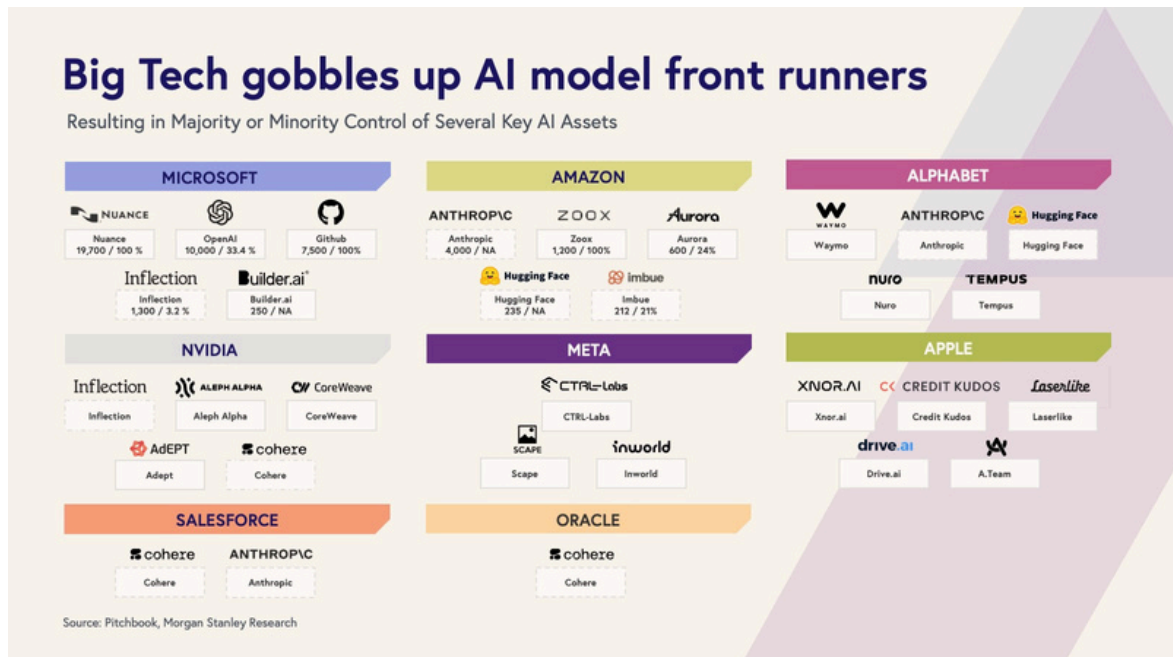
In 2023, foundational model companies captured the lion's share of venture funding in AI, accounting for over 60% of total AI dollars. Players such as [OpenAI](#), [Anthropic](#), [Mistral](#), [Cohere](#), among others raised \$23 billion at an aggregate market cap of \$124 billion, underscoring their critical role in the AI ecosystem globally. Notably, this influx of capital was primarily driven by corporate VCs, who represented 90% of private GenAI fundraising in 2023 (up from 40% in 2022 according to Morgan Stanley). Big Tech companies such as [Microsoft](#), [Google](#), [Amazon](#), [NVIDIA](#), and [Oracle](#), now have significant stakes in foundational model companies, as these investments are strategically aligned to enhance the AI capabilities of these tech giants, driving consumption of their core cloud and compute services. This is in addition to Big Tech companies working on their own in-house foundation model initiatives, such as [Google's Gemini](#) and [Meta AI's Llama](#).

With so much funding flowing into this fundamental layer, competition is intensifying at an unprecedented pace, driving an incredible amount of innovation in the ecosystem. Here are some key trends we've observed in 2023:

- Base models improving rapidly: General purpose LLMs are getting better and better every second, both in terms of base performance capabilities (such as accuracy and latency) but also at the frontier including multi-modal functionality. The launch of [GPT-4o](#) left all our jaws dropping — the new release demonstrated the capability to view and understand video and audio from an uploaded file, as well as generate short videos. The dizzying pace of model improvement raised obvious questions around investment strategy in models that seem to have a half life measured in months.



- Battle between open and closed source intensifies: As we touched upon in last year's State of the Cloud, the open source vs. closed source debate remains a hot topic in 2024 as open-source leaders closely track close-source model performance, especially with the recent launch of Llama 3. New questions have been raised around regulatory impact, whether closed-source players should open-source their older models as part of a new commercialization strategy, or if this might be the first time in history where an open-source leader might become the winner of this market.
- Small model movement gets big: Additionally, 2023 also saw the rise of the small model movement, with [HuggingFace](#) CEO and Co-founder [Clem Delangue](#) declaring 2024 will be the year of SLMs. Compared to larger counterparts, examples like Mistral 8x22b which launched this year have shown that bigger isn't always better in terms of performance, and that small models can have significant cost and latency benefits.
- Emergence of novel architectures and special purpose foundation models: In 2023, we also saw excitement around the emergence of novel model architectures beyond the transformer, such as state-space models and geometric deep learning, pushing the frontier on foundation models that can be less computationally intensive, able to handle longer context, or exhibit structured reasoning. We also saw an explosion of teams training specific-purpose models for code generation, biology, video, image, speech, robotics, music, physics, brainwaves, etc., adding another vector of diversity into the model layer. We discuss this trend in our recent [AI Infrastructure roadmap](#).



With so much happening at the foundational layer, it often seems like the ground is shifting beneath our feet every second. Despite the copious amount of funding that has been invested here, there isn't necessarily a clear consensus on the winner right now.

Prediction: The battle of AI models will remain white-hot for the foreseeable future since this is a critical "land grab" that determines which Big Tech companies reign supreme within the cloud and compute markets in the coming years.

There are three possible realities we expect to see in the foreseeable future on who will capture the most value in this model layer fight:



Reality 1: The model layer becomes commoditized.

Will hundreds of millions of dollars of capital be squandered as VCs and Big Tech back the derby of AI leaders? The most well-capitalized models does not mean they'll become the winner, as open-source models continue to closely challenge the leading market players. But a future where AI models are commoditized doesn't necessarily imply the value of models will diminish. AI models as commodities will be akin to compute or oil as commodities — they'll one day become the assets essential for global business operations. In this reality, the ultimate value in the AI ecosystem will be captured by compute and cloud service providers, marketplaces, and applications — not by the models themselves. However, in a world where AI models are commoditized, as we've seen in the oil market, this could still give rise to one or two extremely valuable companies selling these "commodities."

Reality 2: AI Model Giants split the pie.

Similar to the Cloud Wars, a handful of notable new model companies, heavily supported by Big Tech strategics or corporate VCs, will own the foundational model ecosystem and become giants. Each of these winners will find a differentiated wedge to pair with technological differentiation, whether that's via distribution, price/cost efficiencies, regulatory impact, etc. There could still be a long-tail of different players (especially open-source) but value will accrue to the top handful of model players. It's not only the superior technology that will determine tomorrow's AI Giants, but also their established distribution.

Reality 3: AI models become as diverse and popular as the potato chip market

Just like there are endless flavors of potato chips, the future of the AI model economy could very well look something similar to the snack aisle at your local grocery store. Many model companies can thrive as there are enough differentiated use cases (e.g., modalities, performance, latency, cost, security, etc.) for different model companies to survive. Additionally, geography and regulation could play a role here if geopolitical considerations enter the realm of AI models, with a mix of regulations and sovereign concerns supporting the proliferation of diversity in this layer.

Prediction: While we're far from consensus, a slight majority of our partnership predicts Closed-Source models will drive the bulk of LLM compute cycles, and AI Model Giants will eventually split the economic pie (Reality #2).

We expect to see Cloud Giants leverage their access to compute, chips, and capital to influence the battle in their favor. And the frontrunners are already in the race — Microsoft/OpenAI, AWS/Anthropic, Google/Gemini, with Meta/Llama as the Linux-equivalent OSS alternative, including Mistral as a European lead.

Roadmap: AI Infrastructure

Janelle Teng, Grace Ma, Bhavik Nagda,
Mary D'Onofrio, Elliott Robinson, David Cowan



Venture insights that matter

bvp.com/subscribe

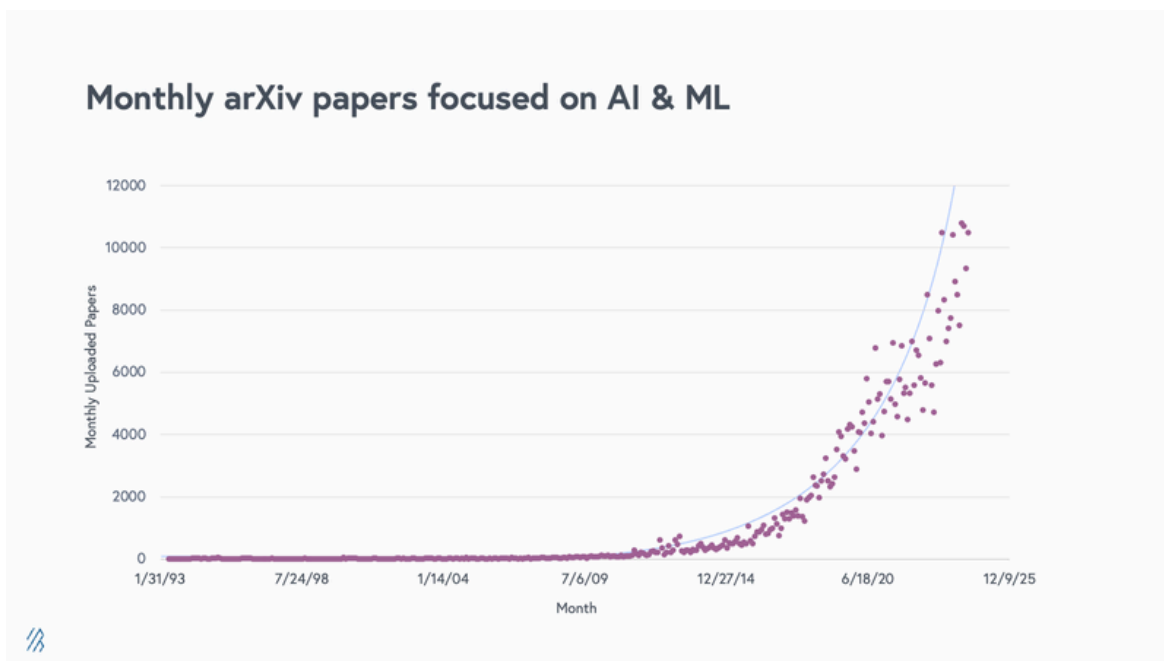


A new infrastructure paradigm, purpose-built for AI, is emerging to supercharge the next wave of enterprise data software in the age of AI.

Bessemer has had a long history of infrastructure investing: From partnering with chip and semiconductor leaders such as [Habana](#) and [Intucell](#), to backing developer platform pioneers [Twilio](#) and [Auth0](#) at the earliest stages, to participating in the modern data stack movement with open-source leaders such as [HashiCorp](#) and [Imply](#). Today, another wave is upon us, with AI ushering a new generation of infrastructure tools purpose-built for enterprises leveraging AI in their platforms.

The AI revolution is catalyzing an evolution in the data stack

Machine learning has dramatically advanced in recent years— since the 2017 breakout paper "[Attention is all you need](#)," which laid the foundation for the transformer deep learning architecture, we have now reached a Cambrian explosion of AI research, with new papers being published every day and compounding at an astonishing pace.



This tectonic shift in AI innovation is catalyzing an evolution in data infrastructure across many vectors.

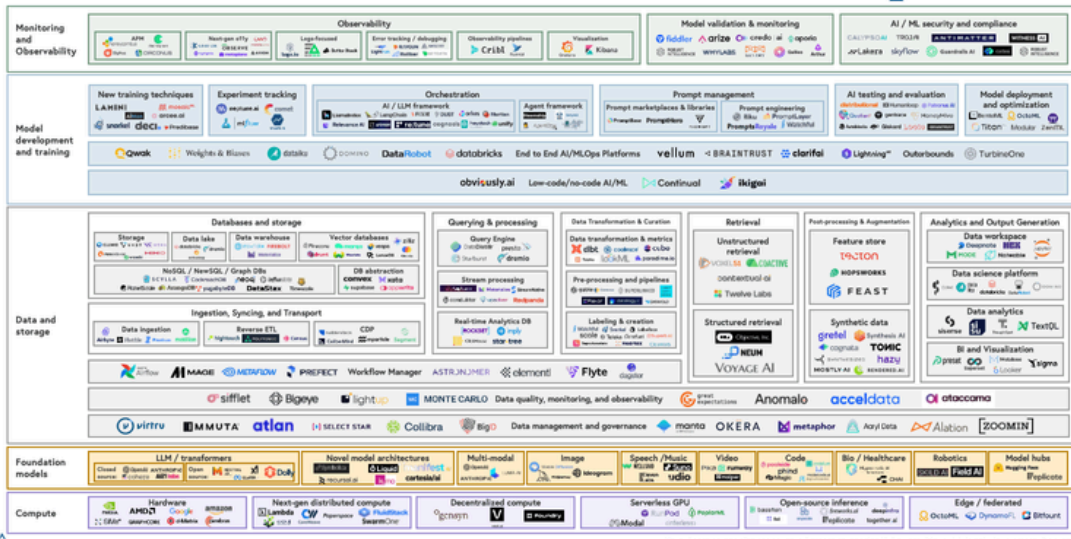
First, AI is powering the [modern data stack](#), and incumbent data infrastructure companies have started incorporating AI functionalities for synthesis, retrieval, and enrichment within data management. Additionally, recognizing the strategic importance of the AI wave as a business opportunity, several incumbents have even released entirely new products to support AI workloads and AI-first users. For instance, many database companies now support embeddings as a data type, either as a new feature or standalone offering.



Next, data and AI are inextricably linked. Data continues to grow at a phenomenal rate to push the limits on current infrastructure tooling. The volume of generated data, especially unstructured data, is projected to skyrocket to 612 zettabytes by 2030, driven by the wave of ML/AI excitement and synthetic data produced by generative models across all modalities. (One zettabyte = one trillion gigabytes or one billion terabytes.) In addition to volume, data types and sources continue to grow in complexity and variety. Companies are responding by developing new hardware including more powerful processors (e.g., GPUs, TPUs), better networking hardware to facilitate efficient data movement, and next-gen storage devices.

Lastly, building on recent progress in ML and hardware, a new wave of AI-native and AI-embedded startups is emerging—these companies either leverage AI/ML from the ground up or use it to augment their existing capabilities. Unfortunately, much of current data infrastructure and tooling is still not optimized for AI use cases. Similar to forcing a square peg into a round hole, AI engineers have had to create workarounds or hacks within their current infrastructure.

The new Data + AI Infrastructure market map



Note: Some companies are repeated across categories if they provide multiple solutions across the stack

With numerous "why now" tailwinds building in recent years, the lack of native and purpose-built tooling has paved the way for a new AI infrastructure stack for AI-native and embedded AI companies.

We are in the midst of a massive technological shift—innovation within this emerging AI infrastructure stack is progressing at an unprecedented pace. Even as we write this roadmap and develop our views, researchers are publishing new papers every day, making previous views obsolete. The rapidly changing environment is intimidating, but the potential and opportunities for startups are expansive, despite unknown variables.

As it often goes, we're investing as the revolution is happening. With new cutting-edge research released daily, it can sometimes feel like the ground is shifting beneath our feet. We are constantly incorporating the latest developments into our thesis. Here are several themes we are drawn to:



1. Innovations in scaling, novel model architectures, and specialized purpose foundation models

The model layer is shaping up to be the most dynamic and hotly contested layers within the AI infrastructure stack. Foundation models are the new "oil" and given the strategic importance of this part of the stack, the winners here may define the future of downstream applications for many years to come as more and more companies build upon their heuristics.

Consequently, we've seen an explosion of activity at the model layer—from open-source to small language models. Much of the activity and capital is focused on scaling transformer-based models (i.e., via data, model parallelism, mixed-modality, etc.) or attempting to push these models across various performance properties (e.g., cost, latency, deployment, memory footprint, context window, etc.). For instance, several teams are improving the building blocks (primitives) of generative models such as attention and convolution mechanisms to create more powerful, capable, and efficient AI technology. Due to the capital intensity of model training, many of these efforts are venture capital-funded. Beyond training costs, a high bar of human capital and specialized resources with the right mix of research and engineering talent are also required to innovate at this layer. We cover more of the current state of innovation, competition, and funding dynamics at the model layer [in our upcoming 2024 State of the Cloud report](#).

But "attention is not all you need"—researchers are developing non-transformer-based architectures, too, and they are continually pushing the limits on what's possible for foundation models. For example, state-space models (SSMs), such as [Mamba](#), and various recurrent architectures are expanding the frontier on foundation models that are less computationally intensive and exhibit lower latency, potentially providing a cheaper and faster alternative to traditional transformers for training and inference. SSMs focused on dynamic, continuous systems have existed since the 1960s, but have only recently been applied to discrete end-to-end sequence modeling. Linear complexity also makes SSMs a great choice for long-context modeling and we're seeing several companies blossom on this front. While early results suggest impressive efficiency across various properties, researchers have a ways to go to demonstrate various properties (e.g. control, alignment, reasoning) now taken for granted in the transformer ecosystem.

Additionally, groundbreaking research within geometric deep learning, including [categorical deep learning](#) and [graph neural networks](#), is equipping researchers with methods of structured reasoning. While this field has existed for quite awhile, it has earned renewed interest in this new wave of AI as geometric methods often enable deep learning algorithms to take into account geometric structures embedded in real-world data (e.g. abstract syntax trees in code, biological pathways, etc) and can be applied to various domains.

Furthermore, beyond general-purpose models, there is currently a proliferation of teams training specific-purpose models for code generation, biology, video, image, speech, robotics, music, physics, brainwaves, etc., adding another vector of diversity and flexibility into the model layer.

2. Innovations in model deployment and inference

The compute layer is one of the most complex layers of the AI infrastructure stack, not only because it's a core layer that quite literally powers other parts of the stack, but it also blends innovations and interactions within hardware (such as GPUs and custom-built hardware), software (such as operating systems, drivers, provisioning tools, frameworks, compilers, and monitoring and management software), and business models. Adding to this complexity, both large incumbents as well as startups are innovating in this area.



At the hardware layer, GPU costs are coming down as supply chain shortages ease. Next-gen GPUs, such as NVIDIA's H100 and B100 series, combined with advancements in interconnect technology, are scaling data and GPU parallelism at the model layer.

Beyond hardware, various algorithmic and infrastructure innovations are enabling new AI capabilities. For example, the self-attention mechanism in the transformer architecture has become a key bottleneck due to its high compute requirements—specifically, quadratic time and space complexity. To address these challenges, the ML systems community has published a variety of model and infra-layer research: evolutions of self-attention (e.g. Ring Attention), KV-cache optimizations (e.g. channel quantization, pruning, approximation), etc. These innovations reduce the memory footprint for the decoding steps of LLMs, unlocking faster inference, longer contexts, and cost efficiencies.

As we move towards personalized, cheaper fine-tuning approaches, many open questions remain. Methods like LoRA have unlocked memory and cost-efficient fine-tuning, but scalably managing GPU resources to serve fine-tuned models has proven difficult (GPU utilization tends to be low as is, and copying weights in and out of memory reduces arithmetic intensity). While improvements in batching, quantization, and higher up the stack in serverless infra have made infrastructure more turnkey, lots of low-hanging fruit remains. Projects like Skypilot and vLLM alongside companies like Modal, Together AI, Fireworks, and Databricks, are pushing the fold here.

Vendors in this layer play an outsized impact on the unit economics (particularly gross margins) of AI application companies that are leveraging their services, and we anticipate these dynamics to continue to drive innovation based on demand from downstream applications.

3. Cutting-edge model training and development techniques

As highlighted earlier, AI research is progressing at a breathtaking pace, and most notably, we are in an exciting period where new AI methods and techniques across pre-training, training, and development, are in bloom. New methods are being developed everyday alongside evolution of existing methods, meaning that the AI infrastructure stack is dynamically being defined and re-defined.

We are seeing these techniques proliferate across all aspects, advancing LLM and diffusion model outputs across base performance parameters (such as accuracy and latency) all the way to pushing the limits on new frontiers (such as reasoning, multimodal, vertical-specific knowledge, and even agentic AI or emergent capabilities). We highlighted a few architectural paradigms in Section I, but other examples of techniques encompass:

- Fine-tuning and alignment: supervised feedback, specialized training data, or refining weights to adapt models for specific tasks (e.g. RLHF, constitutional AI, PEFT)
- Retrieval-augmented generation (RAG): connecting the LLM to external knowledge sources through retrieval mechanisms, combining generative functions with an ability to search and/or incorporate data from relevant knowledge bases
- Prompting paradigms: an interactive process where the LLM is instructed and guided to the desired outcome (e.g. few-shot learning, many-shot in-context learning, step-back prompting, CoT, ToT)
- Model mixing and merging: machine learning approaches that mix separate AI model sub-networks to jointly perform a task (e.g. MoE, SLERP, DARE, TIES, frankenmerging)
- Training stability: decisions around normalization methods (e.g. LayerNorm vs. RMSNorm), normalizations, activations, and other properties can affect training stability and performance
- Parameter efficiency: various methods such as efficient continual pre-training that affect model capabilities and efficiency



While there is a trade-off between simplicity of experimentation versus efficacy of these methods, we predict that these techniques will inspire new developments as researchers iterate faster and solve for real-world scalability and applicability. Furthermore, it is common in applied AI to see a mix or combination of techniques being deployed, but ultimately, the methods that produce the highest bang for buck will likely dominate the applied AI space. Additionally, the landscape is evolving dynamically as base models become better and better and as more AI powered solutions are deployed in production and with real-world constraints.

Ultimately, we believe that we are in early days here and no hegemony has necessarily been established yet, especially for enterprise AI. We are thus excited to partner with companies developing, enabling, or commercializing these techniques as such companies will transform and reimagine how we build, develop, operate, and deploy AI models and apps in reality, and form the key tooling layer for AI companies.

4. DataOps 2.0 in the age of AI

We began this article by claiming that data and AI outputs are inextricably linked. We see this happening across many vectors from data quality affecting AI output (garbage in garbage out), to recent AI innovations unlocking insights from previously untapped data sources (such as unstructured data), to proprietary data serving as a competitive advantage and moat for AI-native companies. We explored this relationship in our [Data Shift Right](#) article, and also highlighted [new data strategies that companies are leveraging to optimize for AI competitive advantage](#) in our recent Data Guide.

Given these catalysts, new demands are being placed on data ops, resulting in the emergence of new approaches and frameworks for storage, labeling, pipelining, preparation, and transformation. A few exciting examples:

- At the pre-processing stage, we are seeing the rise of data curation and ETL solutions purpose-built to manipulate data for a LLM to understand.
- The emergence of new data types (e.g. embeddings) has inspired entirely new data ops categories such as vector databases.
- Data annotation has evolved in the age of AI to incorporate advanced data-centric methods, which have sped up prior manual or weak-supervision approaches, and brought more non-technical end-users into the fold.
- The AI revolution has ushered in the mainstream embrace of tooling for processing various modalities of data, especially unstructured data (such as video and images). Many of these state-of-the-art tools are now integrated into day to day workflows. Previously, dealing with these modalities was challenging and often bespoke, resulting in organizations unable to fully glean value from these rich data sources.
- New enterprise toolchains and data workflows, such as RAG stack, are emerging as organizations leverage innovations in model training and inference techniques (see Section III).

Just as the modern data stack has fueled the rise of iconic decacorns within the data ops space, we believe a new generation of data ops giants will emerge fueled by a focus on AI workflows.

5. Next-gen observability

Alongside each wave of new technology, observability has in turn taken various forms (e.g., data observability in the modern data stack, APM for cloud application development). Similarly, we're seeing observability evolve in the age of AI - a new suite of vendors emerging to help companies monitor model and AI application performance.



While we've seen many companies enter the market solving one key wedge, either in pre-production (e.g., LLM evaluation, testing), in post-production (e.g., monitoring, catching drift and bias, explainability), or even extending into adjacent functions such as model security and compliance, smart routing, and caching, we anticipate (and have already seen) the long-term roadmaps of these companies converging into creating an end-to-end observability platform, creating a single source of truth for model performance in both pre and post-production environments.

We're enthusiastic about a Datadog-like outcome in observability for AI—however, given the ever-changing environment with new models, new training / fine-tuning techniques, and new types of applications, winning in observability will likely involve a team capable of delivering on high product velocity, perhaps more so than in other spaces. As we've gleaned from Datadog's rise, the company was able to break out of a crowded landscape of a dozen or so other (similar) competitors as they focused on a) rapid execution on a broad product and capability set and b) building out deep coverage of what Datadog could monitor, and c) enabling broad integration support so as to bring as many adjacent systems as possible into its ecosystem. We're excited to meet and support this next generation of startups who are taking on such an endeavor for the AI stack.

6. Orchestration

As newcomer LLM and generative AI application companies continue to grow, we see a significant opportunity for companies in the orchestration layer to become the backbone of AI development. With the "orchestra conductor"-like role in the AI development lifecycle and the integral responsibility of ensuring and coordinating the development, deployment, integration, and general management of the AI application, orchestration vendors are a critical (and importantly, vendor neutral) centralized hub that harmonizes the sprawl of various AI tools developers encounter.

Companies like Langchain and LlamaIndex are early breakouts in this space for LLMs, with strong open source ecosystems buoying adoption into companies. They've created frameworks that provide developers with a set of best practices and a toolkit for developing their own LLM applications, abstracting away much of the complexity when it comes to connecting the right data sources to the models, implementing retrieval methods, and beyond. Beyond LLMs, we are seeing an ecosystem of vendors create orchestration solutions for agent-based applications, further streamlining the development process for new innovative agentic AI applications. Much like the success of React in simplifying web development, we anticipate a similar opportunity for AI orchestration vendors to streamline development and enable the masses with the capabilities to develop various types of AI applications (LLM, agent, computer vision, etc).

A massive opportunity exists for AI infrastructure businesses

As Mark Twain once famously said: "When everyone is looking for gold, it's a good time to be in the pick and shovel business." We believe that a massive opportunity exists to build "picks and shovels" for machine learning, and that many multi-billion-dollar companies will be built by equipping enterprises with the tools and infrastructure to operationalize AI.

AI makes us all 10x developer

Lindsey Li, Bhavik Nagda



Venture insights that matter

bvp.com/subscribe

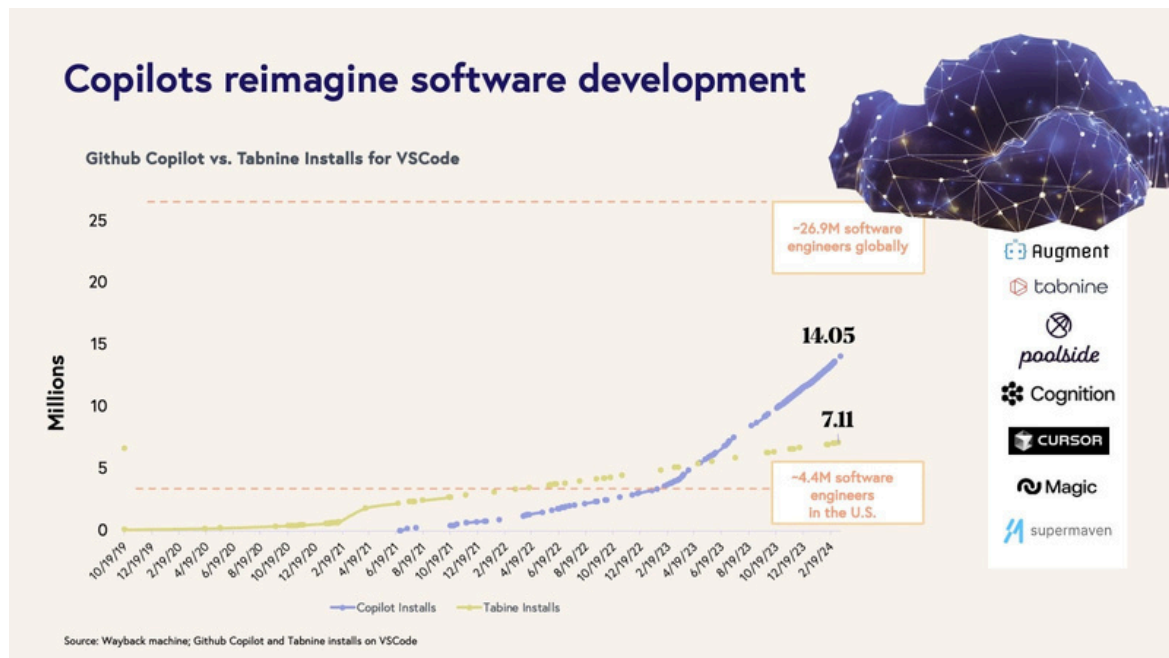


Copilots reimagine global software development

The modern engineer has always been part builder and part student — completing a day job while constantly working to stay up-to-date on new languages, frameworks, infrastructure, etc. The AI quake has added a Ph.D. requirement to the job, as developers face a completely new set of toolchains and best practices for leveraging constantly evolving LLMs, including a new infrastructure suite for data management, curation, prompting, pre-training, and fine-tuning. Each year in the AI era requires coming up to speed on a decade's worth of new developer knowledge.

But AI may also offer the solution to this new level of complexity. 2023 saw widespread adoption of code copilots and the first few months of 2024 saw early breakthroughs in agentic tooling that suggest that end-to-end automation of simple code tasks and perhaps much more may be arriving sooner than we might have expected.

Prediction: The role of the developer will be radically transformed, perhaps more than any other profession, by AI. By the end of the decade, significant developer capability will be available to every human with a computer. The resulting rate of software development will melt keyboards and dramatically reduce the age of the average technology startup founder.



Three main areas are driving the lightning speed evolution of the AI developer economy:

1. The code copilot industry has been a hotbed of innovation and competition with \$3.9 billion VC dollars invested in 2023 for GenAI technology and tools.

Github's incumbent Copilot product, powered by OpenAI's GPT-4 and Codex models, is well-penetrated with north of 14 million installs. A long-tail of well-funded and scaling startup competitors, such as [Tabnine](#), [Magic.dev](#), [Augment](#), [Poolside](#), [Cursor AI](#), [OpenDevin](#), [Cognition's Devin](#), and [Supermaven](#), are building and iterating with developers in the loop. Some, like Magic.dev, Poolside, Augment, and Supermaven, are pre-training their own large AI models with an emphasis on model properties such as context, latency, etc. Others, like Cursor, are model agnostic and are focused on the developer experience, interface, and workflows. This landscape is a good example of the capital intensity of model-layer AI companies; Magic.dev, Augment, Poolside, and Devin have each raised \$150M+ in the last couple of years.



2. The "graduation motion" of copilots embedding agentic search and generation functionality will drive outsized value in the coming years.

Devin, SWE-agent and OpenDevin have demonstrated the potential of end-to-end agentic tools that interact with developer environments (i.e., file editor, bash shell) and the internet to complete coding tasks. Underpinning these agentic demos are rapid advancements in code-language reasoning, agentic trajectory planning (approaches vary across prompting, behavioral cloning / fine-tuning, reinforcement learning), and various agent-computer interface (ACI) improvements (i.e., abstractions and infrastructure across the browser and operating system that enable seamless agentic tool querying and self-correction).

3. Code-language reasoning will remain an epicenter of AI activity, set up to benefit from both model layer innovation (e.g., GPT-4, Claude 3 Opus) and novel reasoning/agentic paradigms (e.g., Cognition's Devin, SWE-agent, OpenDevin).

Model layer improvements will flow down into code editing and completion quality, with value ultimately accruing to developers and software organizations. Beyond code reasoning, systems that push the boundaries of latency, context size, and expand the language domain / pre-training set will also drive outsized value for developers.

AI is driving both innovation and upheaval alike, and accelerating developer velocity, productivity, and leverage for software organizations. Forward-thinking software organizations are routinely surveying the landscape for emerging tools and vendors, and rapidly prioritizing and adopting high-value developer software. Developer budgets are flowing once again and the willingness to pay is high for tools that have visible impact.

For developer entrepreneurs, this is an exciting time to be building; opportunities abound across copilots but also infrastructure, dev tooling, QA, IT configuration and provisioning, security operations monitoring, penetration testing, and on and on.

Copilots are perhaps the most obvious opportunity at the moment, but that makes them likely the most competitive playing field. We have also seen an explosion of tools in more specific developer domains — from SecOps in security to SRE to QA and pen testing. These tools use LLMs to abstract away low-level complexity and automate highly time-consuming, painful engineering tasks, freeing up engineering resources for higher-order tasks. The integration of AI in DevOps processes will enhance CI/CD pipelines, automated testing, and deployment strategies, leading to faster and more reliable software delivery.

Code refactoring is another great example of AI's impact in the developer workflow and ecosystem. Many modern engineering teams spend only a fraction of their FTE time writing net-new code. At large organizations in particular, a large fraction of SWE time is spent on the less "sexy" parts of the software engineering role: maintaining, securing, and testing code. Many of these tasks, such as code refactoring, require deep knowledge of the stack and are often unwieldy projects performed with dread by senior engineers.

AI has obvious potential to address these challenges; startups like [Gitar](#), [Grit](#), [ModelCode](#), and others leverage code-gen models, static analysis, and AST parsers to interpret code structure and migrate code across language, package libraries, and frameworks. Some of these efforts are focused on modern web frameworks while others work with brittle legacy engineering stacks (i.e., COBALT, PEARL, etc.) where fluent engineers are becoming obsolete over time. Many workflows adjacent to the core software engineering function are also highly time-intensive, repetitive, and ripe for automation.



Prediction: By 2030, a majority of corporate software developers will become something more akin to software reviewers. The cost of development will fall and as experienced developers become more productive their salaries will rise.

AI will impact the scope, and skills required for all job markets, but perhaps none more so than of the developer. AI enhancements will not only vastly improve the productivity of this occupation, but also expand the boundaries of the developer universe. By the end of this decade, development capability will be an accessible skill to most of the global population.

The data shift right market

Amit Karp, Lindsey Li, Yael Schiff



Venture insights that matter

bvp.com/subscribe



AI is transforming the data stack. Now, non-technical users are empowered to access business insights more easily than ever before.

Creating and maintaining a "single source of truth" has long been the Holy Grail for data-driven organizations. Getting accurate and up-to-date insights is the crux of everyday decision-making across all business functions. However, the systems that exist today across the data stack fail to make this process easy and accessible — as a result, non-technical users are often struggling to access the right data, when they need it. Defining business metrics, much less accessing the right metric at the right time, shouldn't be so messy and complicated, or involve an entire data team to build tailored solutions. But we're optimistic that data democratization and AI advancements will significantly enhance business users' ability to easily access and leverage data within existing systems while freeing data teams to focus on more impactful problems.

There is a rising crop of new analytics tools that focus on empowering non-technical stakeholders in enterprises. We're calling this the Data Shift Right movement— the professionals across sales, operations, marketing, finance, product and other functions adjacent to technical teams will now become data-fluent and self-sufficient as AI allows users to interact with data in natural language and access the transformation layer without knowing how to code or use SQL.

At Bessemer, we're ready to back founders building in the data shift right space. Here, we explain the common problems founders in the space are setting out to solve, why now is the time to serve this market, and the seven characteristics we think make a best-in-class software solution in the category.

The data problem most leaders can relate to

Today, when business users want to get a better understanding of some of the most relevant KPIs for them, they often work with [data analytics teams](#) to define what they need to measure, how they want a KPI to be presented (e.g. frequency, trends), and where they want it to be delivered. The analytics team will then pull out the data, or create a tailored dashboard for a specific team or professional, which could often take days or weeks (amidst many other priorities), involving back-and-forth communication. This reality leads to inefficiency, larger data teams, and lengthy periods where business users don't have the insights they need. In addition, building these custom dashboards for data analysts is distracting them from larger, more pressing priorities.

Most data-driven organizations are all too familiar with this situation. Tooling that enables business users to become more independent would empower various parts of the organization. Additionally, it would free up data teams to focus on more sophisticated models and research, machine learning, and AI applications, and other strategic initiatives which better serve the needs of the company.

What now? Why now?

In the past few years, the modern data stack has matured significantly and became more standardized across different companies and industries. Changes in technology have opened up new opportunities for businesses to derive value from their data. Data models are now running in the background to optimize companies' core processes, simplify their service, and improve their customers' experience.



Added to this environment, the rapid development and improvement of LLMs and the AI applications built on top of analytics solutions have led to additional game-changing abilities for the "data-shift right" professional:

- The data lineage and transformation layer is now accessible: The transformation layer is where users define the organizational metrics they would like to measure and specify the logic of how these metrics should be calculated. Historically, writing formulas and logic lay in the hands of the data team. This work required (1) strong familiarity with where data is stored, (2) how it flows in the company's data stack, (3) the ability to write the business logic into the right systems, as well as (4) knowledge of SQL to define metrics. But with the advancements in AI, it is now easier to search in your data where certain tables and columns live, understand better the context of the data, ultimately leading to better documentation, improved access, and democratization of the data.
- Data querying is more "natural": In the past, users had to write long SQL queries or use Python to define the data they were looking for, then use BI tools like Tableau or Looker to build dashboards and visualize their data. While many non-technical users learned how to use these BI tools, it wasn't necessarily intuitive for a first-time user and many business users in the organizations would often ask the analytics team to help them pull data rather than doing it themselves. With the availability of AI and LLMs that sit on top of the already established data stack, there has been a new wave of tools that are helping team members across business functions discover and consume data more intuitively, by allowing users to describe in natural language the information they are looking for, and get back raw data, charts, and even simple dashboards. Moreover, users can use natural language to ask for additional changes and edits. Beyond using natural language, users can now interact with data through existing interfaces like Slack, or a simple chatbot, which simplifies the experience even further.

At Bessemer, we believe we are quickly approaching the next wave of data accessibility and visualization companies focused on the business user.

The analytics market has historically been crowded and competitive. With that in mind, we have identified seven characteristics that we think will help tomorrow's data shift right solutions become more successful compared to their competitors and incumbents.

Characteristics of a best-in-class data shift right platform

- Quick time to value: Integrations to existing products should be easy and quick, and ideally don't require a technical team to spend time on implementation. Once the solution is implemented, users immediately get an intuitive product that often includes pre-made dashboards and reports and easy-to-search KPIs.
- Out of the box: Solutions include best practices and benchmarks from leading organizations. Founding teams often arrive with relevant opinions and functional expertise, providing advice on what needs to be measured, how to measure it, and potentially tie the data to organizational processes.
- (Limited) customization: Solutions should allow users to make some adjustments easily — using no code / low code, by providing textual inputs, or by training the underlying LLMs on the business context. At the same time, we acknowledge there is a trade-off between the ease of use and level of customization, and therefore think adjustments should be limited.
- Call to action: Ideally, solutions can not only provide the analysis, but also provide an actionable insight to be taken by the team based on the data. With usage, AI/LLM solutions can "learn" the business context and go on to further recommend fine-tuned insights and next steps unique for the team.



- Source of truth: Tools should be accurate and considered to be the "source of truth" – saving time for other teams and keeping data consistent across the organization. To do so, it's important to keep easy integrations with the existing data stack including BI and other tools used by data analytics teams. Keep in mind that these tools should follow the process of metric and model definition and be able to connect at the model and metric layer so the company doesn't end up with multiple sets of metrics. For example, there shouldn't be a scenario where the sales team and the finance team both define "total sales for the month," but only one of those definitions is available through the company-wide metrics layer.
- Shareable: Findings should be easy to share and able to consume "feedback" – both among core users and teams of the product as well as other internal and external stakeholders. Some common use cases include sharing dashboards over email, creating slides and board materials, or providing QBR materials that are easy to prepare. According to [Operating Advisor Solmaz Shahalizadeh](#), former Head of Data at Shopify, her "dream tool can also 'learn' from the feedback and conversations that happens outside of traditional data tools, about the data."
- Vertical / focused solution: We believe that there is an advantage for companies to focus on a specific business function (e.g. product, marketing operations, sales operations, etc) or alternatively choose a specific business vertical like eCommerce. Having industry or functional context can help tailor the solutions better to their customers needs.

Searching for that single source of truth

Finding that "single source of truth" is a perennial pursuit for most organizations.

Let's look at a specific example with revenue data — it is typically scattered across many systems and is frequently of poor quality. Despite this, RevOps teams depend entirely on this data to make important business decisions like sales rep staffing. Yet, they often have zero access to the data themselves as they often lack the technical skills and access to run SQL queries on the data warehouse. So, a simple question such as, "Which customers have hit their usage cap for the month and should be upsold?" requires a SalesOps leader to submit a request for the data team to run a query, then wait before receiving an answer. The issue only compounds as data analyst teams are pulled in different directions with a long queue of unanswered Jira tickets from across the organization.

The result? Companies miss revenue opportunities when data is siloed in different systems, make poor decisions when using wrong or incomplete information, and waste precious data analyst resources on manual data consolidation and cleanup. To get around this, GTM teams have started manipulating data by circumventing the data warehouse, only exacerbating the problem of data silos and eroding trust in any one system of truth.

There has to be an easier way to make data decisions. Hypergrowth organizations depend on every function learning and iterating their actions based on new insights, without having to depend on technical teams. That's why emerging tools serving Data Shift Right professionals are changing the AI-analytics landscape.

Take these two emergent examples currently in the Bessemer portfolio:

[Preql](#) empowers business users to create and manage their own metrics without relying on a data team. The platform allows non-technical users to connect data sources, develop metrics, and produce reports, facilitating company-wide alignment on key KPI definitions without ongoing manual updates. Preql offers some standard KPIs while allowing customizations to suit individual company needs. Furthermore, using "Preql AI" users can now create custom metrics and dimensions using natural language. This is done while leveraging the customers' existing modern data stack in the background.



Then, there's [Seam AI](#) — a chat interface that can answer any question across all of your customer systems using natural language and AI. Seam's magic is beneath the chat interface — the platform automatically transforms and unifies siloed customer data into one centralized view, enabling teams to uncover new insights and generate powerful customer intelligence. Additionally, Seam can sync these outputs back into customer's business systems to centralize reporting and automate workflows. Seam's models, trained on thousands of queries across the most popular customer systems, produce semantically-correct SQL, ensuring queries are contextually relevant rather than just syntactically accurate.

As Bessemer continues to survey the data shift right market, we've seen many promising AI analytics leaders on the rise to help deliver the Holy Grail companies always seek.

How will AI analytics evolve?

As the analytics ecosystem evolves, several companies are already providing horizontal solutions that cater to general needs rather than specific business functions. These products enhance data retrieval through intuitive interfaces like chatbots or prompts. Simultaneously, traditional BI tools such as Tableau and Looker are integrating AI to simplify their interfaces and improve user experience. We anticipate these solutions will continue to grow momentum as they streamline processes and boost productivity.

Yet, a crucial question remains: who will be the primary users of these tools? (In many cases, users still need at least some technical understanding in order to use these tools.) How will these users be able to quickly generate a business-ready output? We are optimistic about the potential for specialized, vertical solutions to prevail.

Keep in mind that a big part of the data stack for data engineers and analytics will likely remain the same, but will continue to live side-by-side with the data shift right solutions that serve non-technical users. These tools have to talk to each other so they are aligned at the core, but we don't necessarily see the business focused options ever replacing the technical tools.

Accessing data will continue to become an essential need for companies that are looking to improve and make data-driven decisions. As we chart toward a future where data is not just a tool but a catalyst for transformative business decisions, we are excited about a path whereby non-technical users are empowered to access this data themselves for relevant insights and feedback.

Multimodal innovation transforms human relationships with software

Aia Sarycheva, Mike Droesch



Venture insights that matter

bvp.com/subscribe



The rise of multimodal models and AI agents

The rise of multimodal models and AI Agents is leading the next wave of innovation in AI, and dramatically expanding AI's potential applications to far broader use cases than early text-based models achieved. There's a greenfield opportunity for AI entrepreneurs to innovate across new modalities, such as voice, image and video, as well as agentic workflows. These new modalities give AI the equivalent to the human capabilities of vision, hearing and speech, which unlocks the opportunities for AI to play a role in augmenting the large share of human work that is dependent on these senses.

In the next 12 months, we expect voice AI applications in particular to see breakout growth. Over the longer term we also see the promise of agent-first products changing the way businesses operate, as they set new expectations in terms of the complexity and breadth of tasks that AI can be entrusted to handle.

Prediction: Voice AI applications will unlock \$10B of new software TAM over the next five years.

End-to-end voice agents will create massive opportunities as the next wave in voice

Legacy IVR

\$5B+ Market,
poor customer satisfaction,
limited customers/use cases



Modern ASR / Transcription



End-to-End Voice Agents

- Applicable to ALL businesses, not just enterprises
- Applicable to far more use cases
- Potential to replace headcount / do the work
- Very high ROI applications

Recent progress is undeniable

Voice

The first wave of voice AI companies were primarily leveraging advancements in Automatic Speech Recognition (ASR), such as [Abridge](#), which offers the leading product for transcribing notes from doctor-patient conversations, and [Rillavoice](#), which captures field sales representative conversations with customers to assist in sales training.



We are now seeing a new wave of voice AI companies that are developing conversational voice products capable of handling tedious and repetitive workflows, empowering humans across sales, recruiting, customer success, and administrative use cases to focus on higher-value work. One example from our portfolio is [Ada](#), which has taken advantage of recent voice breakthroughs to expand their chat-based customer support product to incorporate conversational voice.

Underpinning these developments are new voice architectures. We are seeing a shift from cascading architectures (ASR transcribes audio to text which is passed into an LLM, then text is fed back into a Text-to-Speech model) to speech-native architectures, as exhibited by new models such as GPT-4o, that can process and reason on raw audio data without ever transcribing to text and respond in native audio. This transition will enable conversational voice products with much lower latency and much greater understanding of non-textual information like emotion, tone, and sentiment, most of which get lost in cascading architectures. These advancements will result in conversational voice experiences that are truly real-time and can help users resolve their issues faster and with far less frustration than prior generations of voice automation.

AI voice applications are emerging in many industries including auto dealerships, retail, restaurants, and home services. A large portion — or even a majority — of inbound sales calls are missed as they happen outside of business hours, and in these cases, AI is primed to pick up the slack. AI voice applications in sales tend to be incredibly high ROI use cases because the AI is essentially picking up lost revenue for these businesses, and can thus offer a really compelling value proposition.

Entrepreneurs building at the forefront of voice AI are more equipped than ever to deliver interfaces that are increasingly natural and conversational, capable of providing near human-level performance. We expect to see an explosion of companies across the Voice AI stack (see below), many of which will experience truly breakout growth. In the process, we also expect consumer expectations around interacting with voice AI to change, as modern conversational voice applications start to deliver far more natural experiences for users and ultimately get them to resolution much faster.

Image / Video

Computer vision models have existed for years, but what is so exciting about the new generation of multi-modal LLMs is their ability to combine their understanding of image and text data (among other modalities), as this combination is extremely useful for many tasks.

The initial wave of enterprise-based image applications was focused largely on data extraction use cases. We have seen companies like [Raft](#) ingest freight documents, extracting critical information to populate the customer's ERP and automate invoice reconciliation workflows. As the underlying models keep improving, we believe we will see a host of vertical-specific image and video processing applications emerge that will also be able to ingest increasing amounts of data to power their applications.

We have also seen applications in engineering and design that leverage vision models, and image generation models to help reason on graphical data, like schematics, or generate renderings of a building design. For example, [Flux.ai](#) offers an AI copilot that helps electrical engineers generate printed circuit board components in their design software, based on ingesting a PDF spec sheet for the component.



Autonomous AI Agents

One of the most exciting emerging themes in AI is the development of AI Agents, capable of handling complex multi-step tasks end-to-end, fully autonomously. While most AI agents don't yet operate reliably enough to function autonomously in complex use cases, progress on agentic workflows is moving very quickly and we are seeing glimpses of what is possible. Each new demo is better than the last, with Cognition AI's Devin — the AI software engineer — hinting at what's possible as AI's planning and reasoning capabilities continue to expand.

More applications are beginning to implement AI agents in highly constrained use cases in which they can limit the impact of compounding errors across multistep processes. For example, enterprises are leveraging solutions like Bessemer portfolio company [Leena AI](#) providing AI agents to support employees with IT, HR, and Finance related tasks, helping these teams free themselves of busywork and improving the employee experience.

In addition, new models are emerging with stronger reasoning capabilities that can further empower agents to execute more complex workflows. And perhaps more interesting, there is a flurry of research focused on new architectural approaches to improve agent implementations through various methods including, chain-of-thought reasoning, self-reflection, tool use, planning, and multi-agent collaboration.

2023 was the year we saw an explosion of AI applications focused on text-based use cases. In 2024, we predict multimodal models will open up new frontiers in terms of both the capabilities and use cases that we see AI being used in at the application layer. This will lead to a new wave of applications bringing near human capabilities to markets ranging from large enterprises to small businesses within specific verticals, and will even unlock exciting potential for consumer apps.

Vertical AI dwarfs legacy vertical SaaS with new applications and business models

Caty Rea, Brian Feinstein



Venture insights that matter

bvp.com/subscribe

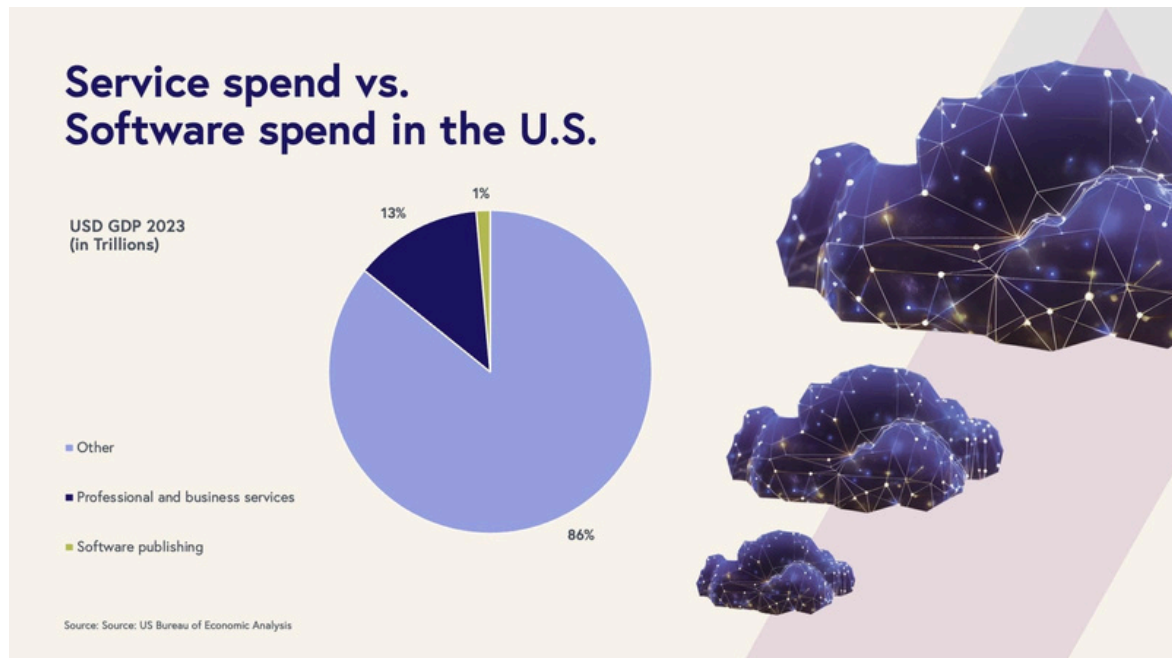


Vertical SaaS proved to be a sleeping giant that transformed industries during the first cloud revolution. Today, the top 20 US publicly traded vertical SaaS companies represent a combined market capitalization of ~\$300 billion, with more than half of these companies having IPO'd in the last ten years.

Now the rise of large language models (LLMs) has sparked the next wave of vertical SaaS as we see the creation of new LLM-native companies targeting new functions and at times industries that were out of bounds for legacy vertical SaaS; notably Vertical AI applications target the high cost repetitive language-based tasks that dominate numerous verticals and large sectors of the economy.

The US Bureau of Labor Statistics cites the Business and Professional Services industry at 13% of US GDP making this sector alone, dominated by repetitive language tasks, ~10x the size of the software industry.

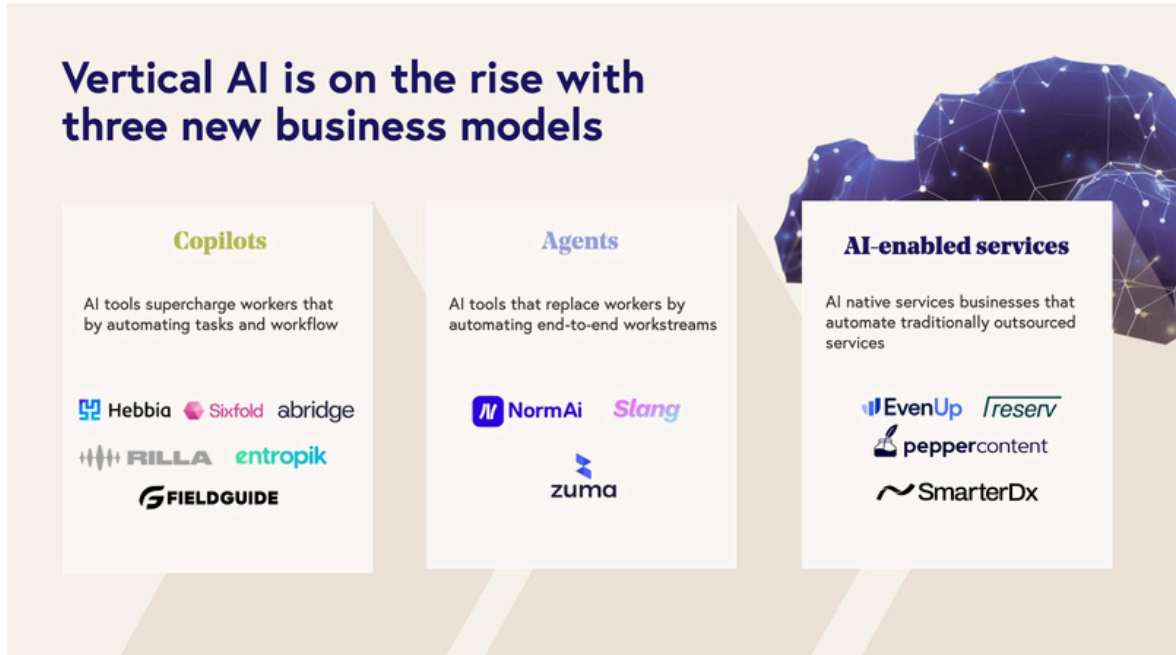
Beyond the professional service sector, within every industry vertical repetitive language based tasks represent a significant share of activity. We believe vertical AI will compete for a meaningful share of these dollars and will also drive activity in areas where human labor was insufficient. For example, Bessemer portfolio company EvenUp automates third party legal services as well as internal paralegal workflows. Moreover, EvenUp opens up task areas where human labor was formerly too expensive or inconsistent to be applied. This multi-dimensional expansion holds implications for Vertical AI across all sectors of the economy.



Prediction: Vertical AI's market capitalization will be at least 10x the size of legacy Vertical SaaS as Vertical AI takes on the services economy and unleashes new business models.



Copilots, Autopilots, and AI-enabled Services make up the three new business models of the Vertical AI economy. Vertical AI is also being delivered via several different business models, thus increasing the odds of matching AI capability with a given industry need.



Copilots accelerate efficiency among workers by leveraging LLMs to automate tasks. Sixfold, for example, supercharges insurance underwriters to better analyze data and understand risk. In the copilot model, the AI application sits side-by-side with the human user to make the user more successful.

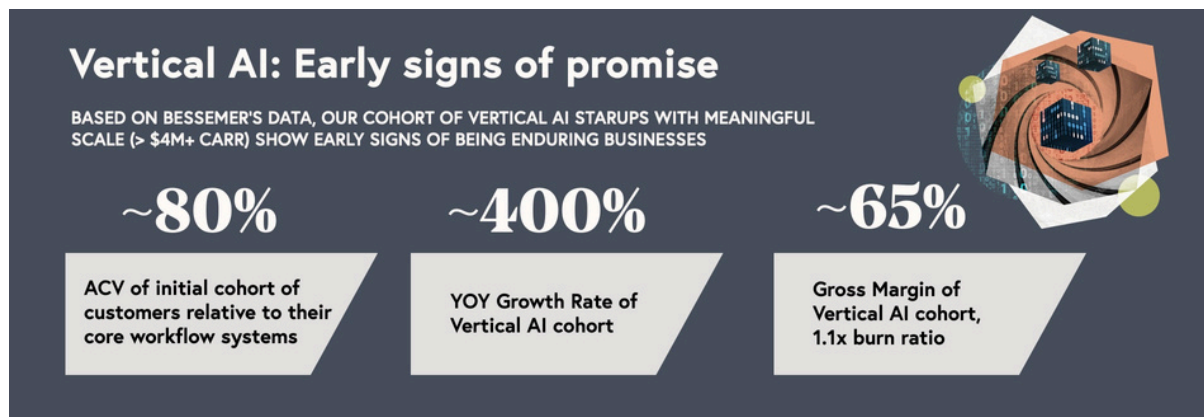
While copilots help employees do their work, Agents fully automate workflows and replace the user. Agents tend to focus on specific functions inside of vertical companies like outbound sales or inbound call reception. Slang AI, for example, handles inbound calls for restaurants to book and reservations, answer questions, and more.

Finally, we are seeing the emergence of AI-enabled Services. These are services typically outsourced to a third-party provider like accounting, legal services, medical billing, etc. Because they are so people-intensive, these businesses have traditionally been lower margin, hard to scale, difficult to differentiate, and less valuable than technology businesses. By using software to automate work, these AI-powered services companies aim to deliver cheaper, better, and faster services to the market and take share from incumbent service-oriented businesses. SmarterDx, uses AI to audit inpatient claims on behalf of health systems and hospitals before the bill and corresponding clinical documentation are sent to a payer. These pre-bill services were traditionally outsourced to vendors that used physicians and nurses to do this audit work.



Early signal on Vertical AI business model strength from the Bessemer portfolio

Bessemer was fortunate to back the legacy SaaS leader in several verticals — and now we have one of the largest Vertical AI portfolios, particularly with businesses that have reached mid to growth stages. As a result, we already have meaningful data we can use to compare Vertical AI companies and legacy vertical SaaS comps. And while we're as excited as any VC about the power of language models, we're growing equally excited by the early data we're seeing on Vertical AI business models. Three analyses of our Vertical AI portfolio hint at the strength of this new class of applications.



First, we'll note that for the most part Vertical AI players are leading with functionality that is not competing with legacy SaaS. The utility of these applications is typically complementary to a legacy SaaS product (if one exits at all) and thus not being asked to replicate and displace an incumbent. Equally exciting, these Vertical AI upstarts are already commanding ~80% of the ACV of the traditional core vertical SaaS systems. And these Vertical AI players are just getting started with obvious potential to expand ACVs. This data demonstrates Vertical AI's ability by replacing service spending to unlock significant spend within vertical end markets and deliver TAMs that may ultimately be a significant multiple of those enjoyed by legacy SaaS.

We're also encouraged by the efficiency and growth profile of our Vertical AI companies with meaningful scale (\$4M ARR+). This is a cohort of companies growing as fast as any we've ever seen at ~400% year-over-year. As impressive, these companies are also demonstrating healthy efficiency with an average ~65% gross margin and a ~1.1x BVP Efficiency Ratio (Net New CARR / Net Burn). We believe these companies will only improve margins over time as we've historically seen in software, and thus are, as a category, well positioned to be enduring standalone public companies.

Finally, we analyzed the percent of revenue these Vertical AI companies are spending on model costs to address the concern that many of these applications are simply thin wrappers. On average, these companies are currently only spending ~10% of their revenue on model costs or ~25% of their total COGS. Thus these vertical applications built on top of LLMs are already generating margins ~6X the underlying model expenses. With model costs dropping rapidly and these startups just starting to optimize their spend, we believe these attractive margins will only get better. Overall while we expect massive value creation in the model layer, this data tells us that as with past infrastructure innovations the majority of enterprise value will once again be captured in the application layer.

Vertical Software incumbents are not completely asleep at the wheel. Companies like [Thomson Reuters](#) (acquiring [CaseText](#) for \$650M) and [DocuSign](#) (acquiring [Lexion](#) for \$165M) have made some of the first high-profile Vertical AI acquisitions.



But we believe we're still near the starting line for a Vertical AI marathon...albeit one where the runners may sprint the entire race. With early startup leaders such as EvenUp, Abridge, Rilla, Axion, and others growing at impressive clips, we expect new enduring public Vertical AI companies to be born in a few short years.

Based on the growth rates at scale we are already seeing, we predict we will see at least five Vertical AI Centaurs (\$100M+ ARR) emerge within the next two to three years.

Prediction: The first Vertical AI IPO will occur within the next three years.

AI brings Consumer Cloud back

Maha Malik, Talia Goldberg, Kent Bennett



Venture insights that matter

bvp.com/subscribe



It's no secret that the consumer cloud has had a slow decade. (We define consumer cloud as companies that provide cloud-based storage, compute, and digital applications directly to individual consumers, including, at times, concurrent B2B and "prosumer" offerings.)

To illustrate just how slow, we analyzed the past eight years of [Cloud 100 data](#)—the definitive ranking of the top 100 private cloud companies published by Bessemer, Forbes, and Salesforce Ventures every year since 2016. Only 4% of the cumulative lists since inception nine years ago represented companies with a consumer offering, sometimes alongside a much more prominent B2B offering (e.g., Zoom in 2016 and, more recently, OpenAI in 2023). Arguably, we have not seen an exit of a 'pure' consumer cloud company since the once-upon-a-time decacorn, Dropbox, which had their IPO in 2018.

Eight years of the Cloud 100

We have not seen a major consumer cloud exit in five years.

	2016	2017	2018	2019	2020	2021	2022	2023
Cloud # 1	slack	stripe	stripe	stripe	snowflake	stripe	stripe	OpenAI
Cloud # 100	Skyhigh	Cansva	pendo	DASHLANE	LaunchDarkly	AXONIUS	Figma	Deepl
List Value at C100	\$99B	\$116B	\$138B	\$166B	\$267B	\$518B	\$738B	\$654B
# of Consumer Cloud companies in C100	3	4	3	2	3	4	5	9
Consumer Cloud companies in C100	Dropbox eventbrite zoom	Dropbox eventbrite zoom Cansva	eventbrite zoom Cansva	Cansva DASHLANE	Cansva grammarly Notion	Cansva grammarly Notion Calendly	Cansva grammarly Notion Calendly iPassword	Cansva ANTHROPIC OpenAI iPassword brightwheel
# Exits of C100 Consumer	3	3	2	--	--	--	--	--
Exit Value of C100 Consumer	\$28B	\$28B	\$19B	--	--	--	--	--

Consumer cloud unicorns have historically been built in the aftermath of major enabling technology shifts. But we haven't seen a widespread relevant quake in consumer facing technology since the launch of the iPhone and subsequent developments in social media platforms almost fifteen years ago. However, two years ago consumers heard a major rumble.

As the fast-evolving multi-modal capabilities of LLMs allow us to extend and enhance our text, visual, and auditory senses in previously impossible ways, we're seeing potential for disruption open up in every category of legacy consumer cloud.

One measure of AI's consumptive power is how much these applications gobble up our time and attention. For example, ChatGPT is now neck-to-neck with leaders in the Attention Economy, such as Reddit, with other general-purpose AI assistants, including Claude and Gemini, quickly gaining traction.

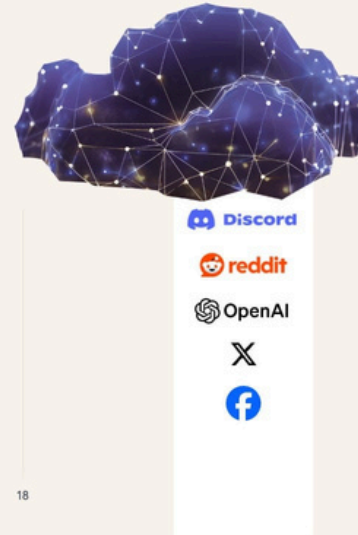


AI takes on the Attention Economy

MONTHLY SITE VISITS
(BILLIONS, MAY 24)



Source: Similarweb



In addition to general-purpose assistants mentioned above, we're already seeing examples of consumer AI companies that are driving innovation in their respective categories. These include [Perplexity](#) for search, [Character.ai](#) for companionship, [Midjourney](#) for image creativity, [Suno](#) and [Udio](#) for music generation, and [Luma](#), [Viggle](#) and [Pika](#) for video generation. These companies are demonstrating the potential of LLM-native applications to attract and retain a dedicated user base and in some cases, effectively displace modern incumbents.

With AI changing the way we engage and play with technology, this is one of the most exciting times to be a consumer cloud builder and investor. We expect multiple consumer cloud IPOs over the course of the next five years.

Prediction: With the startling rise of synthetic media, new consumer applications, and conversation AI-agents, we predict that by 2030 the top three businesses dominating the Attention Economy will be based on AI-generated content or products.

We're also seeing significant early stage activity in the longer tail of functionally specific consumer AI applications (i.e., content generation and editing, education) as evidenced by monthly web and app visits. The good news is that these signals indicate the amount of consumer demand and excitement – an early indication that consumers are looking for AI to enhance their lives. The bad news is that there are not yet more than 10 category-specific consumer AI-native apps that have shown clear signs of product depth beyond thin wrappers or proven sustaining customer love in the form of strong retention. We believe there is still a clear opportunity for motivated entrepreneurs to build sustaining cloud companies over the coming months and years to address many unmet consumer needs.



As we look across consumer needs, we're asking ourselves two key questions to understand where value will accrue in this LLM moment:

- How acutely painful or labor intensive is the status quo for the consumer?
- How much repetitive, predictable language / visual / auditory effort is required?

As we ask these questions, we are reexamining every daily need and pain point in a consumer's life, but also not limiting our imagination on what's possible by just considering established consumer needs. We believe there are large businesses to be built delivering novel utility to consumers, such as clones and companions, creativity and creation, interactive entertainment, and memory augmentation, including many other yet-to-be-invented markets.

We are also excitedly tracking novel form factors that are starting to surface to address specific consumer needs. Since we can't predict the future, we can't tell exactly what form factor AI will take as it penetrates consumer life. However, hand-held devices, wearables, and household objects (toys, frames, mirrors) are already starting to emerge, at least as prototypes, as potential harbingers of startups to come.

AI will not only reinvent our favorite pastimes (e.g., social, entertainment, shopping, travel, etc.), but also help us discover and reimagine new ways for people to connect, play, buy, and explore the world.

There is plenty yet to be figured out. From an investment standpoint, we question which consumer demands will be fulfilled by general-purpose AI assistants (including, for example, Siri on the iPhone) versus standalone applications. Not to mention ethical considerations which will emerge alongside these future products. Despite many unknowns on the horizon, we believe the early signals clearly indicate that the LLM revolution will change all of our lives, and rejuvenate the consumer cloud landscape.

The autonomous robotics future is around the corner

Alex Ferrara, Aditya Nidmarti, Bhavik Nagda



Venture insights that matter

bvp.com/subscribe



A world where robots can learn, act, and help humans autonomously has been years in the making.

Intuitive Surgical's robots helped perform 1.6 million surgeries last year; often procedures that would have been impossible for a surgeon alone to perform safely. That's a staggering number until you consider the fact that it accounts for less than 0.5% of surgeries performed globally. Robotics is an exciting field that holds the potential to unlock substantial value for society, but it's still a field that is arguably in its infancy.

Today, vertical markets like manufacturing, logistics, and medicine deploy robots to sharpen and hasten their workflows—but they require either human control or programming to follow pre-defined rules in order to operate. Think of pick-and-place robots in factories, or even Intuitive's DaVinci doctor-controlled surgical robot. These robots excel at replacing repetitive human tasks, improving output while saving time and costs.

In the last several years, we've seen the very early adoption of autonomous robots—robots that can perceive their surroundings and act without human input. Autonomous robotics is still a niche field, but we believe that its emergence will expand the entire robotics market in the coming years as it enables higher-value use cases.

In the wake of artificial intelligence's (AI) watershed 2023, we believe autonomous robotics is set to have its moment soon, too. Developers are applying AI technology, like multi-modal models and large language models (LLMs), to solving challenging robotics problems, such as perception, path planning, and motion control. In this deep dive, we explain why now is the time to start paying attention to autonomous robotics, including recent market developments and growth opportunities ahead.

Why is autonomous robotics relevant now?

Historically, robotics development required significant investments in time and capital just to build prototypes. But recent technological advancements, such as generative artificial intelligence (AI) and neural networks, have accelerated key phases in the development timeline. Today, robotics has one of the highest NASA Technological Readiness Levels (TRL) across deep tech—near-final products are passing tests and approaching commercialization. (Think Cruise's driverless car service in San Francisco, which became publicly available in 2022.) Now, many are asking the question, "Can autonomous robotics companies become great businesses?"

Unlike software, it takes a lot of sophisticated inputs to train autonomous robots—more than code alone. Autonomous robotics learn how to operate in two primary ways: through cameras and computer vision, and deep learning and generative AI.

Building blocks of autonomous robotics

1. Cameras and computer vision

To perceive their environment, autonomous robots rely on cameras and computer vision technology. In the future we think autonomous robots will use a range of sensors including Lidar, Radar, Sonar, and other wavelengths of the electromagnetic spectrum to perceive the world in a way that is impossible for humans. But most of today's sensors other than a camera are cost-prohibitive for most applications. High quality cameras, however, have become inexpensive thanks to 20 years of usage in smartphones.



Smartphone adoption, and the associated mass production of cameras, has driven substantial improvements in the performance per dollar of cameras in a manner that has not occurred with other sensors. Cameras typically also use passive sensor technology, which is less expensive to manufacture than active sensors such as LiDAR or RADAR. Active sensors typically bounce a wavelength of light off a target and measure the time it takes for the beam to return to the sensor. This often requires mechanical components, such as the rotating LiDAR sensor seen atop Google's autonomous vehicles, which are more expensive to manufacture at scale than solid state devices.

Why now?
Cheap cameras and computer vision

CAMERAS →

Millions of units sold per annum

2021

1,400.0
1,300.0
1,200.0
1,100.0
1,000.0
900.0
800.0
700.0
600.0
500.0
400.0
300.0
200.0
100.0

■ Film cameras (negligible)
■ Digital cameras (Blue)
■ Smartphones (1.4B) (Yellow)

The rest is history

Intel® RealSense™ Depth Camera
D435iif
\$354.00

Equipped with cameras to help them see, robots just need an extra push to help them process what they're seeing. Computer vision technology allows robots to segment objects in one plane and track them as they move across different planes. In other words, it helps them recognize what they're seeing over time and understand the world as humans do.

Why now?
Cheap cameras and computer vision

CAMERAS → **COMPUTER VISION**

OBJECT DETECTION **IMAGE SEGMENTATION** **DEEP TRACKING**

OBJECT DETECTION

IMAGE SEGMENTATION

DEEP TRACKING

Current frame Search Region Core Layers Fully Connected Layers

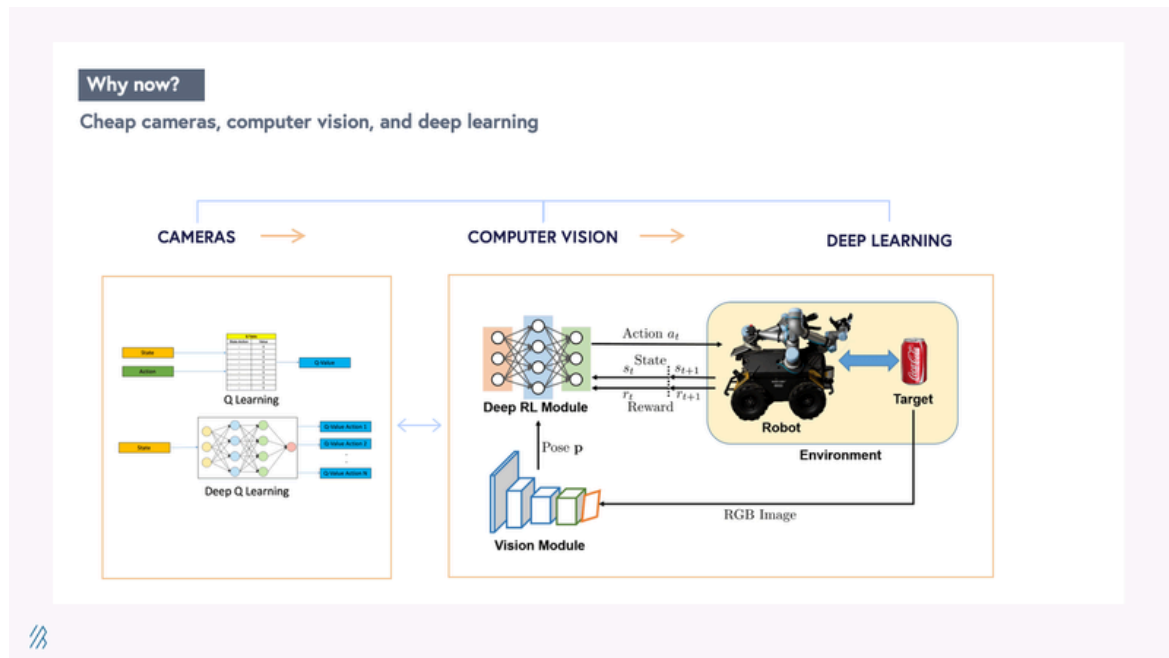
Previous frame What to track Core Layers Predicted location of target within search region



2. Deep learning and generative AI

To gain autonomy, robots need to be able to learn over time, and eventually gain the ability to make decisions on their own. Deep learning, which is enabled by neural networks, allows robots to make sense of what they're seeing and learn from the commands they're given. Furthermore, recent innovations such as Google DeepMind's Robotic Transformer 2 (RT-2) now enable a vision-language-action paradigm whereby robots can ingest camera images, interpret the objects in a scene, and directly predict actions for the robot to perform. Each task requires understanding visual-semantic concepts and the ability to perform robotic control to operate on these concepts, such as "chop the celery" or "clean up the mess". This has the potential to reduce the friction between the human-machine interface by allowing humans to direct autonomous robots using natural language rather than esoteric robot programming languages.

With technological maturity in cameras, computer vision, and deep learning, we get a robot that can perceive, understand, and learn from its environment; a robot that's autonomous. Generative AI removes the friction for human-robot interactions, making communication as easy as speaking to a colleague or friend.

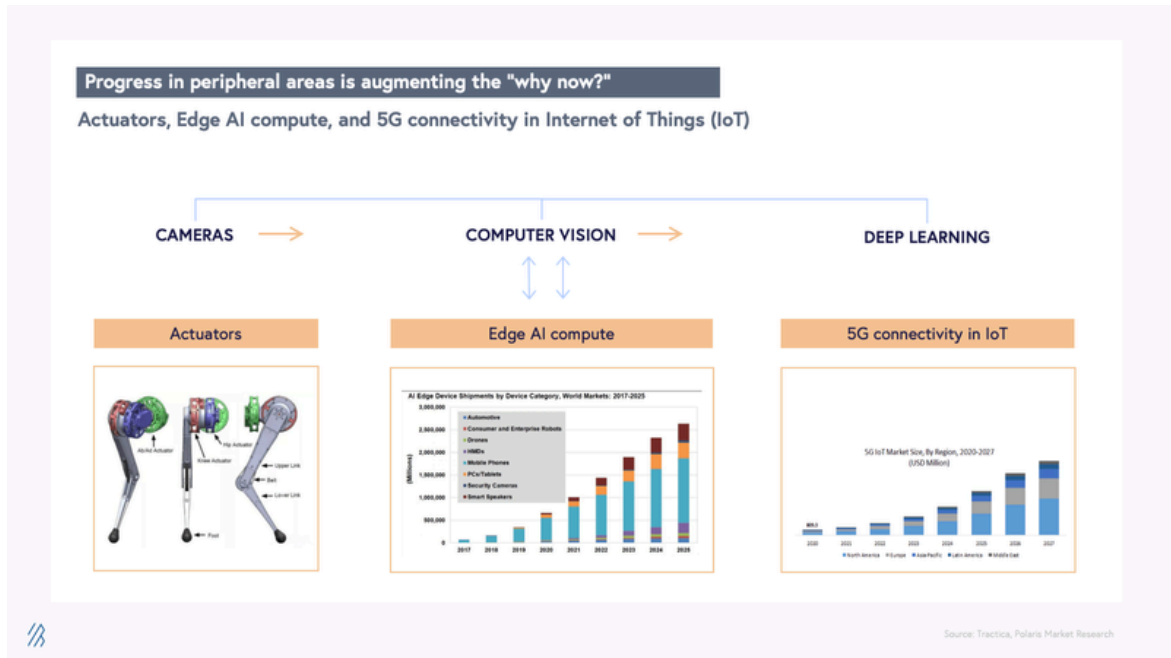


Progress in a number of peripheral areas has also accelerated innovation where autonomous robotics is concerned:

- **Actuators:** Historically one of the weak links for robot design, actuators must be safe, durable, power and cost efficient while also generating high levels of torque output and density. The dexterous hardware acts much like a human joint, giving robots the ability to move fluidly. This is one of the weak links for robot design because actuators must be safe, durable, power and cost efficient while also generating high levels of torque output and density. Companies such as ANYbotics are working on some exciting innovations in this area.
- **Edge AI compute:** Advances in edge computing means that robots will be able to process more powerful computations at the source, delivering quicker and more secure results compared to cloud-based robotic systems where data is sent elsewhere and analyzed. This enables a swath of real-time and high-security use cases that weren't previously possible.



- 5G connectivity: Fleets of robots can connect to and learn from one another using 5G, even in areas where WiFi signals are weak. Fleet learning enables learning at scale to unlock an exponential increase in robotic capabilities. Advances in eSIM/iSIM standards and chip availability mean 5G cellular connectivity will become a viable option for many use cases where power usage had previously made it impractical.



Our prediction: autonomous robotics is on its way to help humanity

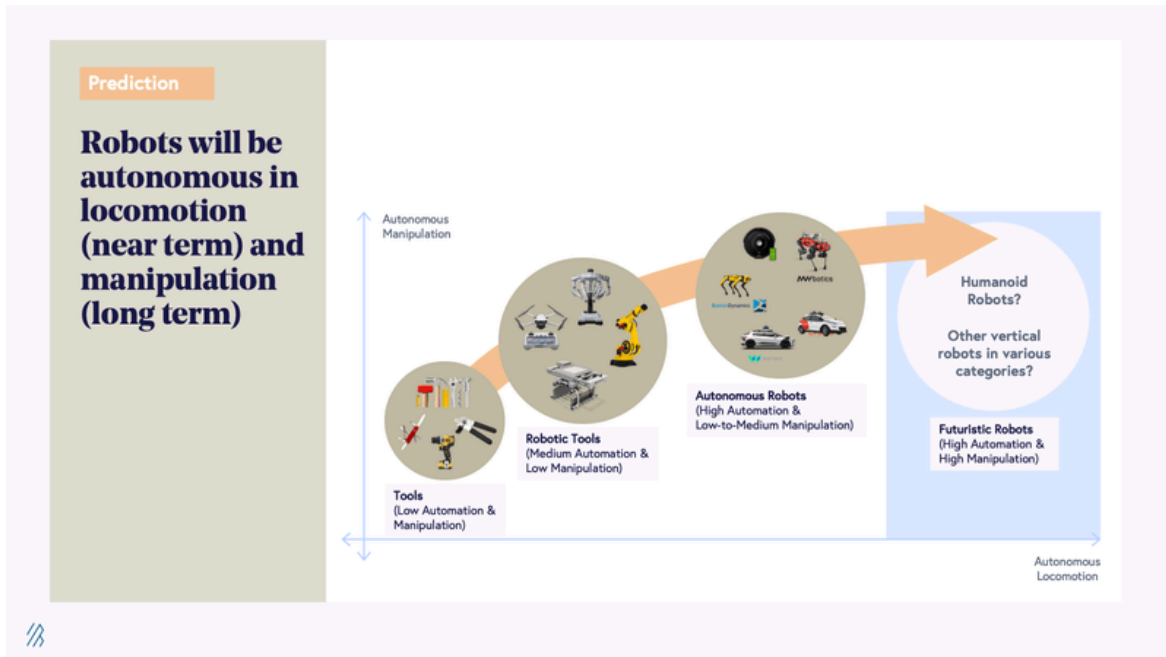
When we talk about autonomous robotics, there are two ways in which we see that autonomy emerge: autonomous locomotion and autonomous manipulation.

Autonomous locomotion enables a robot to go from point A to point B on its own (think of the Cruise example from earlier). Conversely, autonomous manipulation is where we see the potential for robots that not only impact productivity, but actually protect — and in some cases enhance — human life.

This kind of autonomy enables robots to do human-like, dexterous tasks on their own, such as pushing things, opening doors, or taking samples.

Having robots perform tasks in high risk environments, such as oil rigs and nuclear plants, would take humans out of inherently dangerous work environments, not to mention preventing a number of tragic casualties. Companies like Figure, 1X Robotics, and Tesla are already building and testing humanoid robots that can perform mundane and unsafe tasks.

As companies seek to increase robots' autonomy and rely less on remote operation, product developers will need to train models on egocentric datasets—visual data from the first-person perspective. New AR/VR products like the Apple Vision Pro offer the potential to accelerate the creation of this training data.



We predict that autonomous and remote manipulation will change a number of industries, including:

Medical robotics and automation

Medical robots assist surgeons with non-invasive surgeries, improving surgical processes, surgeon ergonomics, and patient outcomes. In the future, artificial intelligence (AI) and machine learning (ML) will improve surgical performance and patient outcomes. One day we can expect to see these robots perform surgeries on their own, using deep learning to execute procedures well.

Warehouse and fulfillment robotics and automation

E-commerce fulfillment situations, especially in warehouses, accelerate fulfillment with the use of robotics. With the addition of 3-D vision and 5G enablement, robotics will be able to take more more complex tasks within decision-heavy scenarios.

Construction, cleaning, and inspection

Robots are leveraged to inspect work sites (using aerial drones), automate repetitive processes, aid in welding; injection; and finishing, and help clean work sites (Roomba and ECOVACS come to mind).

As robots get smarter, they'll be deployed in hazardous industrial environments where humans can't safely go, and can even guard them with security automation. They'll also contribute to efficiency improvements with tasks like painting, coating, and inspection of machinery.



Human companionship

This one might feel a bit more sci-fi than the others, but it's also one of the most exciting future use cases to imagine. As robots gain intelligence and dexterity, they could be available for elderly and disabled people who need help living independently. Having a robot aid in meal preparation, cleaning, and personal hygiene — and even friendship — would make a huge difference for those that are struggling.

Military drones

Uncrewed aerial vehicles (AEVs) can provide militaries with a competitive advantage in warfare. Companies like [Anduril](#) are not only building autonomous drones for military use, but also autonomous vehicles that can intercept and destroy other drones.

Hurdles to overcome

Of course, just like any emerging field, the development of robotics has seen its fair share of roadblocks and challenges preventing wider commercialization. We're actively looking for companies that are well-versed in the following challenges, and looking for innovative ways to overcome them.

- **Hardware-related execution challenges:** Robotics technology needs hardware, and hardware is at the mercy of a long supply chain. Coordinating these (quite literally) moving parts increases complexity once production is outsourced. Executing this well is extremely important, and it's also difficult to pull off.
- **Capital intensity:** Robotics companies need inventory. This results in inventory build-up and RaaS contracts where the company has to hold the robot as an asset on their balance sheet. In a world of higher interest rates, financing working capital becomes more expensive.
- **Revenue quality:** Most robotics companies start off by selling hardware for one-time revenue, making it difficult to find opportunities for recurring revenue. These companies haven't found a hardware and software combination that would result in better revenue. Furthermore, businesses that have a mix of recurring and one-time revenue are often more difficult for the investor community to assess since traditional SaaS KPIs may not be applicable.
- **Employment:** Robotic automation disrupts global employment. On average, the arrival of one industrial robot in a local labor market coincides with an employment drop of 5.6 workers. This will apply pressure to policy makers, so we're asking companies to think of how regulation might impact their vision from day one.

Our guiding principles on how we will invest

We're looking to back strong technical teams that have deep robotics engineering and supply chain expertise building autonomous robots that enable tasks which were previously impossible. This is our top criteria. Beyond this, we're interested in startups that aspire to build businesses that can achieve the following:

- **Disrupt Large markets:** Targeting large Total Addressable Markets (TAMs) when calculated on a bottoms-up basis.
- **Enable complex use cases:** We're not interested in labor arbitrage (low level) use cases; there's already plenty of activity in these markets. Instead, we're looking for use cases that at the very least protect human life. Ultimately, we want to move in the direction of enabling things that humans previously couldn't physically or technically achieve (like our oil rig example above).
- **Connected, cloud-based fleets:** Remote telemetry for fleet learning is a particular feature we're on the lookout for. Equipped with remote telemetry, robots can transmit data to other nearby robots on the fly, who can then take that data and learn from it. We're also looking for companies that are actively thinking about autonomous manipulation.



- Recurring revenue and high average revenue per account (ARPA): A portion of revenue must be recurring, and average selling price (ASP) must be six figures, with tons of room available to increase actual cash value (ACV) and ARPA. As a directional example, think more Intuitive Surgical and less Roomba. Even if penetration is low, these metrics would signal that the room to penetrate is much higher.
- Traction forward is a bonus: We're looking for companies with robots in active use in production environments for a minimum of 6 months and avoiding firms where the sales have solely been to innovation groups.

The road ahead with autonomous robotics

The technology behind autonomous robotics is closer than ever to widespread deployment. From medical and warehouse use cases to construction and companionship, we expect to see autonomous robotics touch a diverse set of industries. Also on the horizon is teleoperation—robotics manufacturers and customers have the opportunity to remotely operate robots, and in some cases can assign the work to people in other parts of the world. ANYbotics, for instance, has been experimenting with teleoperation since 2019, when it announced one of its products could perform industrial inspection tasks, like checking energy plants for rust and leakages, via teleoperation. When Elon Musk shared a video of Tesla's humanoid robot Optimus folding a shirt, he explained it was remotely operated. Phantom Auto, too, built a platform that enables factory workers to remotely operate forklifts. Teleoperation represents a meaningful intermediate step between today's human-controlled, in-person robots, and fully autonomous robots. Even with all the data and use cases we currently have, we can't make a definitive prediction about where autonomous robotics is headed. But with the right investments and the right resourcing, it's not hard to imagine how further advancements will make a staggering number of "what if?" scenarios a reality.

Unlocking machine learning for drug discovery

Andrew Hedin



Venture insights that matter

bvp.com/subscribe



Advances in biology and medicine over the last decade have given us some of the most transformative medicines to date, including the first FDA-approved gene therapy, Luxterna, to treat retinal dystrophy, the first engineered cell therapy, Kymriah, to treat lymphoblastic leukemia, and multiple vaccines to treat the worldwide COVID-19 pandemic in less than a year since the virus was first discovered.

In 2020, the industry set a new NASDAQ biotech record by raising nearly \$16 billion across more than 80 IPOs. By the second month of 2021, biotech IPOs had already raised nearly \$3 billion cumulatively.

Despite these scientific breakthroughs and voracious investment appetite, the \$1 trillion global pharmaceutical industry has been experiencing a productivity slide dating back to the 1950s. In just the last ten years, large cap biopharma companies have seen returns on investment in drug discovery decline from 10.1% to 1.8% and forecasted peak sales per drug cut in half. Pharmaceutical giants now often prefer deploying their cash reserves to partner with or acquire biotechnology startups versus investing heavily into internal R&D efforts; a third of forecasted sales are derived from acquisitions and a quarter from co-developed drugs. Many large cap biopharma companies are also facing the looming patent expirations on historical blockbuster drugs, leaving them actively searching for biotech startups and drug discovery methods to diversify their drug pipelines.

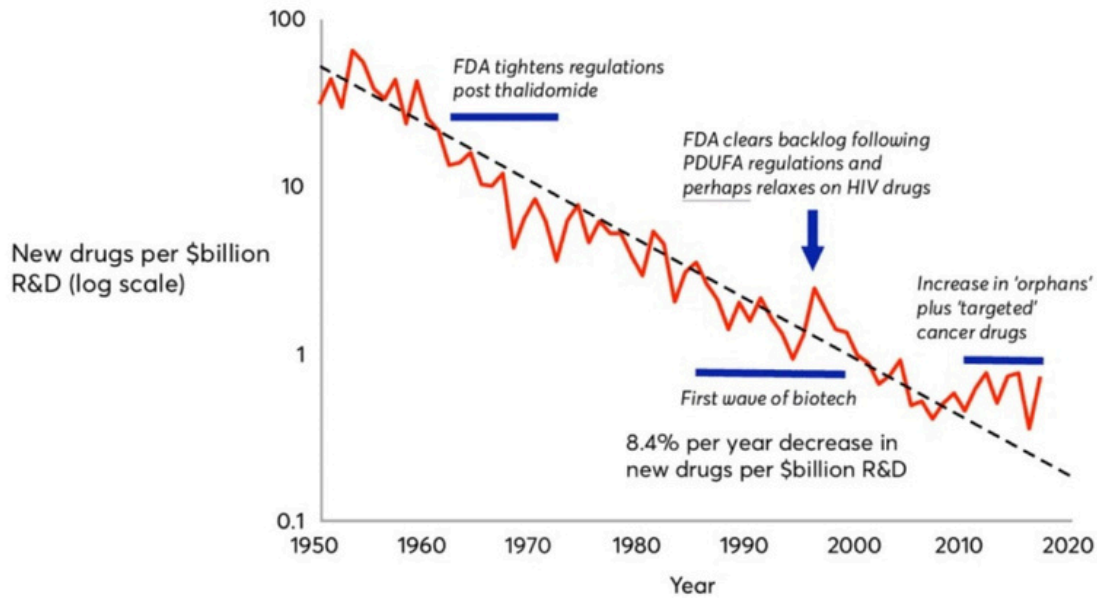
Over the last decade, Big Pharma has been betting on machine learning to usher in a new era of faster, cheaper, more-efficient, and advanced drug development. Despite some pharma industry concern that hype has exceeded reality in AI/ML drug discovery to date, we are now seeing the fruits of the pioneering work early evangelists have poured into the field. Most pharmaceutical giants have announced a partnership with a machine learning company, many firms are building machine learning R&D teams in-house, and the first drugs discovered via machine learning have now reached early-stage clinical trials.

We've previously written about how software is making the clinical trials process more efficient and remain interested in companies across the pharma IT value chain. Here, we share our perspective and roadmap on the entrepreneurial opportunities for machine learning in preclinical drug discovery engines.

The steadily rising cost of drug discovery

Moore's Law, coined in 1965 by Intel co-founder Gordon Moore, predicted that computing power would double every 18 months. In drug development, we've witnessed the inverse. Pharmaceutical companies are spending increasingly more to develop fewer drugs. Eroom's Law (Moore spelled backwards) states that the number of new drugs approved per billion U.S. dollars spent on R&D has halved roughly every nine years since 1950, an 80-fold drop in inflation-adjusted terms. Although estimates are hotly debated, today, of the five percent of drug candidates that make it to market, the development process costs roughly \$2.5 billion.

That \$2.5 billion is composed not only of the direct R&D costs for the commercialized drug, but also the aggregate costs of all the failed drug candidates a drug company endured over the journey and the opportunity cost of investing their liquid assets elsewhere.



Why is drug discovery so expensive?

It's in part inherent to the nature of biology. We still don't understand biology well enough to build robust, bottom-up models of what proteins to target or how drugging a target of interest will affect an individual cell, much less an organ system or the entire human body. Most drug targets are parts of complex cellular networks leading to unpredictable changes (i.e. drug side effects). Biological systems also show a high degree of redundancy, which could blunt the effects of even the most specifically-targeted therapeutics.

It's also inherent to the nature of the market. Commercially-viable drugs have to be significantly better than currently available treatments, so as more drugs are approved, building a new drug for a disease with prior treatments becomes increasingly more difficult. This explains why many pharma companies have focused heavily on rare diseases and tumors with unique mutations (known as precision oncology), where a new drug may be the first and only treatment available.

The last two decades have also experienced bias towards brute force approaches to drug discovery. The industry has shifted from iterative medicinal chemistry coupled with phenotypic assays to the serial filtering of a static compound library against a given target. As a result, we've overestimated the ability of high-throughput screening of large chemical libraries to a specific target to tell researchers that a drug candidate will be safe and effective in human clinical trials.

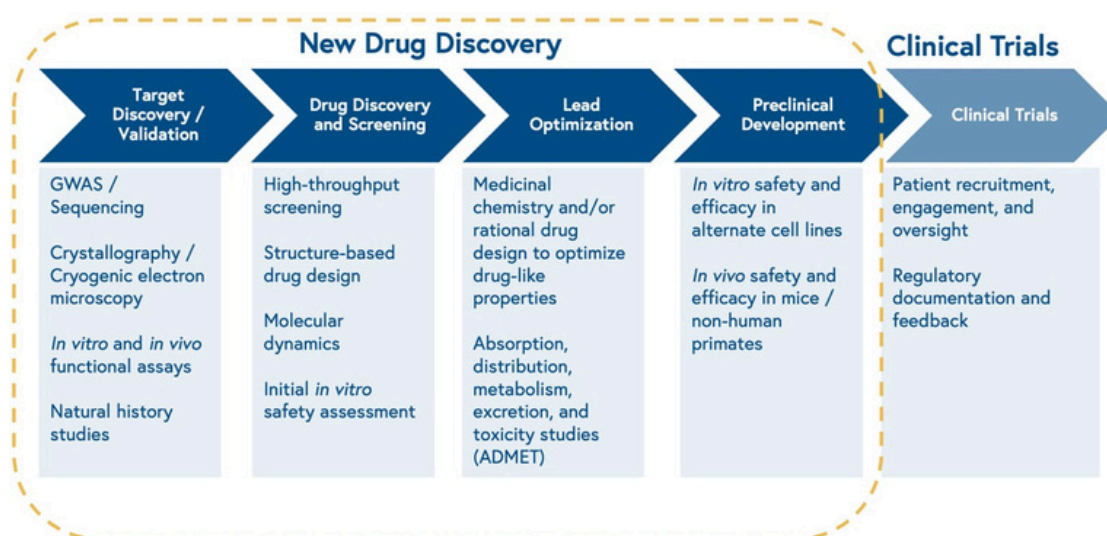
In fact, [more first-in-class small-molecule drugs in the first half of the 2000s were discovered using old-school phenotypic assays versus today's drug-target affinity assays, even though the latter has been the new industry standard for the past two decades](#). Phenotypic screens, although slow, capture and abstract away the complex network of unpredictable effects from drugging a target, which is impossible to identify in a lone drug-target binding affinity assay. Capturing, representing, and perturbing that complex biological network in silico is the holy grail of machine learning for drug discovery.



Deconvoluting the drug discovery process

Drug discovery, the process of identifying novel therapeutics to treat disease, is the first component of the biopharmaceutical value chain. We've covered ways how software can make the latter components of the biopharmaceutical value chain, including clinical trials, more efficient [elsewhere](#). The drug discovery process may differ based on whether a drug candidate is a small molecule or a biologic (e.g. proteins, antibodies, cell therapies, gene therapies). The majority of ML approaches to drug discovery to date have focused on small molecule development, although recent companies have emerged to tackle biologics.

The discovery process begins with studying the biology of a cellular target, i.e. something a drug can 'hit'. Once the biology is well characterized, the process continues with high throughput screening of thousands of existing chemical compounds to find those that 'hit' the target of interest, known as leads. Next, scientists make iterative chemical changes to lead molecules to maximize properties of interest, such as binding affinity to the target, and test the distribution, clearance, and toxicity of the drug in a petri dish and animal models.



This approach is slow, limited by the size of physical chemical compound libraries, and prone to high attrition rates. Consequently, only one out of 10,000 molecules screened for a given target will make it through the drug development process.

Why now?

The pharmaceutical industry is riding tailwinds from scientific revolutions in both biomedicine and computer science. There are three drivers highlighting why now is the key time to focus on the potential of machine learning in drug discovery.

1. New tools have improved our ability to create, manipulate, and measure biological systems at scale.

Advances in molecular biology and bioengineering over the past decade have improved our ability to represent human biological systems in smaller sizes and test therapeutic hypotheses at scale. This begins with our ability to create models that more accurately reflect unique biological dynamics and variation, including induced pluripotent stem cell (iPSCs)-based models of different cell types, organoids that reflect three-dimensional tissue dynamics, and patient-derived xenografts (PDX) to mimic human disease in mouse models.



Next is our ability to perturb these models in a genome or protein-targeted fashion. The development of multiplexed CRISPR/Cas9-based gene editing gave researchers the power to quickly insert, delete, modify, or replace multiple genes in living organisms.

CRISPR screens allow scientists to activate or knock out the function of one or many specific genes at once in high-throughput and identify genetic dependencies unique to certain cell types (known as synthetic lethality). Directed evolution of proteins allows scientists to mutate structures to yield advantageous characteristics, such as tissue or biomarker-specific targeting and evasion of the immune system.

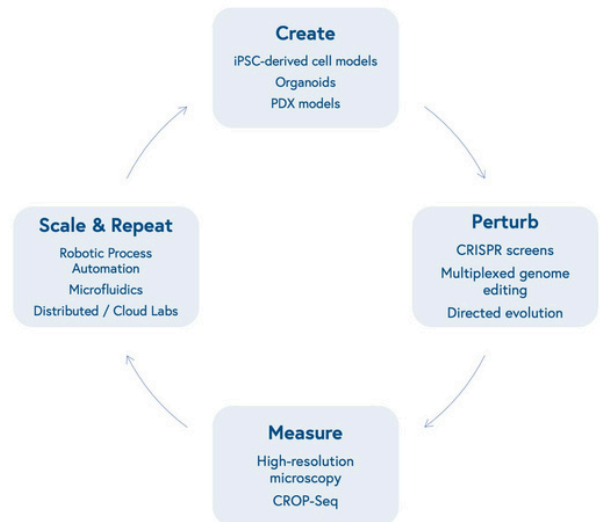
Subsequently is our ability to measure those perturbations at incredibly high-resolution. This can be accomplished through advances in microscopic imaging, including the ability to visualize live-cell dynamics, or more granular genetic sequencing, including the ability to measure changes in gene expression level at the single-cell level (single-cell RNA sequencing) after perturbations with CRISPR screens (CRISPR droplet sequencing, or CROP-Seq).

Finally is our capacity to scale and repeat this cyclical process. Robotic automation and microfluidics within in-house labs or via virtual, distributed labs, including contract research organizations and virtual labs like Strateos, Emerald, and Culture Biosciences allow biotech entrepreneurs to outsource and scale otherwise manual, tedious, and repetitive biological experiments, making it easier and cheaper to produce proprietary biological data. These data may offer new insights into biological systems, providing researchers with the knowledge to create new models that reflect varying diseases and repeat the loop.

2. Machine learning models are delivering better representations of biological systems

In the realm of software, the cloud and machine learning infrastructure that has been a cornerstone in SaaS and redefined dozens of industries is making its way to biopharma.

Newer model architectures, such as gated graph sequence neural networks and quantum machine learning, are making progress towards better characterizing the spatial, electrical, and energetic states of drug candidates, their biological targets, and the high-dimensional biological systems they interact within.



Fueling those representations are larger and more granular biomedical, chemical, and clinical data collection and distribution efforts on the part of the government, hospitals, universities, and biotech firms. Most notable of these data collection efforts is the National Institute of Health's 'All of Us' research program, launched in 2018, which is actively collecting longitudinal health and genetic data from 1 million volunteers to help scientists make progress in precision medicine. A notable challenge with publicly aggregated data sources (e.g. PubChem, ChEMBL, TCGA, ImmPort) is that the raw datasets are often incomplete, noisy, unstructured, and poorly understood, contributing to well-known 'garbage in, garbage out' concerns. The last decade, however, has benefited from a surge in open-source bioinformatics tools to clean datasets for functional use as inputs into statistical machine learning models.

Made possible by the advances in biology mentioned earlier, both large biopharma firms and young startups are building end-to-end approaches to training machine learning algorithms, generating proprietary, high-quality data at the lab bench to feed into algorithms to build better predictive models.



3. Bilingual talent is applying computing power to ask the right questions of complex biological systems

A new breed of entrepreneurs, including bilingual teams that closely pair engineers with biochemists, and scientists trained in both biology and computer science, are paving the way for machine learning to ask the most compelling questions of the biological data we have on hand.

The partnership between scientists, machine learning engineers, and clinicians not only asks what problems are worth solving, but also which problems can we solve and what data do we need?

For all its promise, machine learning in biomedicine will depend on human talent to set research directions, generate novel biological questions and hypotheses, produce and filter data, and validate results. We're likely to see hybrid models with bespoke assays run by scientists and standard experiments run by machines and software.

Opportunities for machine learning in drug discovery

Machine learning applies algorithms to learn from data and then either characterizes or makes predictions about new data sets. Three factors influence the potential of machine learning to make useful prediction in drug discovery: 1) the specific features of a molecule, target, or biological system available as input data, 2) the quality and quantity of data available, and 3) selection of the right model architecture for the task at hand.

Machine learning algorithms can take a variety of biological features as inputs, including genetic sequencing data, small molecule structure libraries, biochemical assay data, microscopy images, or text from academic literature. These models can largely be broken up into either supervised or unsupervised learning. Supervised learning aims to predict future outputs of data, such as regression and classification, whereas unsupervised learning aims to identify novel hidden patterns or relationships within high-dimensional datasets and cluster similar data together. There are opportunities for both in the drug discovery value chain.

Drug discovery value chain



1. High-throughput in silico analyses of multidimensional data could identify novel biological targets



Drug discovery relies on the development of drug candidates, including small molecules, peptides, antibodies (and more recently nucleic acids and cells) designed to alter disease states by modulating the activity of a molecular target of interest. Identifying a target of interest relies on a therapeutic hypothesis, i.e. that modulating a given target will lead to a change in the disease state.



Novel biological measurement tools, including high-dimensional microscopy and single-cell RNA sequencing are well-suited as multidimensional training data for unsupervised machine learning models to infer new, meaningful relationships between biological components, i.e. potential targets, and disease.

We have encountered two (not mutually exclusive) classes of companies applying machine learning methods in this space:

1. Precision medicine (i.e. -omics-based) startups sifting through a combination of public and self-generated genomic, transcriptomic, and proteomic data to identify and validate novel targets (e.g. [Insitro](#), [Octant](#), [Deep Genomics](#), [Verge Genomics](#)). Within this class are two additional approaches: Combination target discovery: identification of genomic and metabolomic correlates of response to standard-of-care treatments, such as immune checkpoint inhibitors (e.g. [Ikena Oncology](#), [Immunai](#)). Synthetic lethality: identification of paired genomic targets that, when perturbed or mutated in tandem, lead to cell death (e.g. [Tango Therapeutics](#), [KSQ Therapeutics](#), [Artios Pharma](#)).
2. Imaging-based startups applying convolutional neural network architectures to high-resolution digital microscopy to evaluate phenotypic cellular changes when perturbed with a drug candidate of interest. As mentioned earlier, this high-throughput phenotypic approach captures and abstracts the complex network of unpredictable effects (i.e. polypharmacology) that scientists may otherwise miss in modern drug-target affinity assays that evaluate one compound against one target at a time. (e.g. [Recursion Pharmaceuticals](#), [Eikon Therapeutics](#)).

Furthermore, as modern biomedical science becomes increasingly rich in data, products like [Watershed Informatics](#) are commoditizing bioinformatics and machine learning tools to enable both academic biologists and pharmaceutical companies to quickly run initial target discovery analyses on raw sequencing data.

Finally, biomedical literature is the primary setting of our collective knowledge of associations between molecular targets and disease. Companies like [nference](#) are processing unstructured text with natural language processing (NLP) to identify relevant papers and associations between diseases, targets, and drugs among the trove of digitized literature. This approach has been particularly useful for identifying new applications of previously commercialized drugs or shelved drug candidates whose data have been described largely in free text.

2. Model architectures that represent biochemical structures, properties, and interactions with potential targets could identify more promising drug candidates



Typical small molecule drug discovery involves experimental, laboratory-automated, high-throughput screening of millions of diverse chemical compounds against a protein target of interest to identify 'hits'. Hits are then biochemically modified to optimize target specificity, selectivity, and binding affinity.

Representing these three-dimensional structures and thermodynamic interactions between drugs and targets *in silico* is both mathematically and computationally challenging. Laboratory approaches to identify and visualize protein structures, including X-ray crystallography and cryogenic electron microscopy are expensive and difficult, and, even when a structural representation of a molecule is available, simulating protein movement/dynamics is computationally demanding.



In the 1980s, the biopharma industry developed early versions of statistical and biophysical modeling programs for molecular docking prediction. These were limited by both limited computing power and an incomplete understanding of how molecules interact. New unsupervised learning methods and novel model architectures to represent molecules, such as graph convolutional networks, can derive their own insights about which biochemical features matter to develop better predictions of protein structure and drug-target interactions.

The most notable recent computational development in structural biology came late last year, when DeepMind's [AlphaFold2](#) algorithm was able to reliably predict over 90% of a single protein structure based on the amino acid sequence alone. While this feat is not at the level of resolution needed for target discovery yet, it remains a promising advance for the future of protein structure prediction. More immediately, technologies like AlphaFold2 may be useful for designing protein-based therapeutics, including antibodies and peptides, where high-resolution prediction is less necessary.

Startups today are leveraging new model architectures, vast public and proprietary biochemical structure and assay data, and novel biological tools to identify more promising hits, i.e. machine learning can give us better starting molecules to move through the drug development process. We've categorized companies tackling this problem into three groups:

1. Deep learning-guided simulations of molecular docking, i.e. the spatial interaction and binding affinity between a potential drug and the target protein. (e.g. [Atomwise](#), [BenevolentAI](#), [Nimbus Therapeutics](#)). Within this class are two additional approaches: Structural Allostery: Applying computational biophysics, genomics, and molecular dynamics for identification of novel, non-competitive (allosteric), binding sites on target molecules (e.g. [Relay Therapeutics](#), [Hotspot Therapeutics](#), [Frontier Medicines](#)). DNA-encoded Libraries (DELs): allow biopharma companies to inexpensively predict drug activity from up to 40 trillion molecular structures, stored in a single test tube mixture, by linking distinct DNA sequences as unique barcodes for each tested biochemical compound. Emerging biotech companies are using these data to train machine learning models to predict drug activity on more diverse compounds that weren't included in the DELs (e.g. [ZebiAI](#), [Anagenex](#)).
2. Molecular de novo design via generative model architectures, including recurrent neural networks, variational autoencoders, and generative adversarial networks, have been used to create small molecules and protein therapeutics with optimal bioactivity, pharmacokinetics, and other desired properties (e.g. [Insilico Medicine](#), [Generate Biomedicines](#)).
3. Repurposing or identifying synergistic combinations of off-patent drugs, which by nature of commercial use have de-risked safety profiles, to modulate novel targets (e.g. [Pharnext](#), [BioXcel Therapeutics](#)).

3. *In silico* prediction and optimization of pharmacodynamics / pharmacokinetics via better in vitro microphysiological systems could improve the success rate of drug candidates in clinical trials.



Lead optimization involves rational chemical modification and evaluation of lead compounds for desired properties, including pharmacodynamics (the effect of the drug on living organisms) and pharmacokinetics (how the living organism acts on the drug) both in silico and in vitro. Pharmacokinetics is better described by the acronym ADMET: absorption, distribution, metabolism, excretion, and toxicity (i.e. off-target effects). For each lead, this process involves synthesizing and testing a large number of similar molecules in silico and in vitro.



In silico prediction of undesirable off-target effects of leads and associated idiosyncratic toxicity could minimize the number of eventual failures of drug candidates in human clinical trials. This has long been a challenge that Big Pharma has gone after, and companies like [Pfizer](#), [Bayer](#), [Sanofi](#), and [Bristol-Myers Squibb](#) have all published their in silico approaches to lead optimization and ADMET prediction. Improved in vitro microphysiological systems, such as three-dimensional organoids and organ-on-a-chip devices, offer a higher quality and more biologically-relevant data source for predictive models compared to traditional petri dish assays.

Startups in this space are both evaluating public assay data to predict optimal chemical modifications for lead candidates and building high quality in-house bioactivity and ADMET datasets themselves (e.g. [Reverie Labs](#), [Genesis Therapeutics](#)).

4. Computational prediction of biomarkers correlated to drug response in vitro and in vivo may improve the success rate of drug candidates in human clinical trials.



Optimized compounds are finally tested in an in vitro model of disease, from cell lines to organoids, to determine how well the drug candidate is modulating the target of interest within a living system. Safety and efficacy of the drug candidate is then tested in varying animal and patient-derived xenograft (PDX) models, ranging from mice to non-human primates, depending on the target and disease of interest. However, this process is expensive, time-consuming, and only weakly correlated to how well the drug will fare in humans. Some startups were built to serve biopharma companies faster than traditional contract research organizations with automated collection of animal model data (e.g. [Vium](#), acquired by [Recursion](#)).

To help improve the success rates of clinical trials, biotech companies will often employ unsupervised learning to identify translational biomarkers in these preclinical models that predict better response to a drug candidate. This has been particularly useful in the case of precision oncology and immunotherapy, in which genomic signatures have predicted drug sensitivity in a tissue-agnostic manner. Identifying predictive biomarkers can improve selectivity in clinical trial recruitment and stratify patients in later-stage trials (e.g. [Scorpion Therapeutics](#), [Volastra Therapeutics](#)). Additionally, companies like [PathAI](#) and [Paige](#) are supporting biopharma with computational pathology as a service, identifying spatial biomarkers that correlate to drug response.

New frontiers

Developing in silico models that deliver realistic, meaningful representations of life remains a major challenge in biology and represents a large opportunity for scientists and engineers. Two emerging areas of interest for machine learning we're interested in are in de novo protein engineering and quantum mechanics-informed simulations of molecular dynamics.

Protein engineering through machine-guided directed evolution enables scientists to optimize desired protein features through the creation of novel variants. These methods predict how protein sequences map to function without requiring prior knowledge of the underlying protein structure, molecular dynamics, or associated biological pathways.

A key trend in biopharma has been the growth of biologics, i.e. protein-based therapeutics such as antibodies and enzyme replacement therapy, due to lower attrition rates, strong safety profiles, and defensibility from competing biosimilars relative to small molecules, which more easily lose market share to generics.



Companies like Manifold Bio, LabGenius, Serotiny, and Nabla Bio are working on machine learning-based approaches to protein engineering. AbCellera and BigHat Biosciences are applying these same principles directly to the development of more selective and less immunogenic antibody therapeutics. Finally, Dyno Therapeutics has carved a niche in designing bespoke, machine intelligence-informed adenovirus-associated vectors (AAVs) to preferentially deliver gene therapies to specific organs and avoid provoking an immune response to the viral vector itself, which has troubled gene therapies of years past.

As computing power continues to scale in the cloud year over year, companies like Silicon Therapeutics (acquired by Roivant Sciences) and ProteinQure are developing quantum mechanics engines to explore interactions beyond drug-target binding, including loop dynamics and domain breathing, which may more accurately represent specific, complex conformations of targets and identify unique binding sites for drug candidates in silico.

The response to the COVID-19 pandemic by both software and biopharma is a testament to the resilience and ingenuity of both industries, and we believe there is no better time than now to see their core strengths converge to develop the next-generation of lifesaving therapeutics. We are excited to continue partnering with and learning from the scientists and engineers working on novel computational approaches to design better medicines faster.

Lessons for AI leaders



Venture insights that matter

bvp.com/subscribe

Six imperatives for AI-first companies

Morgan Cheatham and Steve Kraus



Venture insights that matter

bvp.com/subscribe



Change happens slowly, and then all at once — especially in complex industries like healthcare. Just five years ago, venture capital [investments in healthcare AI were emerging and exploratory](#). Half a decade and one global pandemic later, we're living in a brave and more ambitious new world defined by an unbridled enthusiasm for leveraging revolutionary technologies like AI. Pointing this technology at previously intractable problems in key industries such as healthcare, life sciences, and beyond is among the greatest opportunities of the century.

2022 was the year the broader public bore witness to material advancements in AI research that have matured from lab to life. ChatGPT educated over 100 million people globally about transformers in just two months. What was once a nascent area of research has now become technology's next platform shift, and with that, investors ask — how will generational AI companies be built in healthcare, life sciences, and beyond?

AI-first companies are in the business of advancing AI as a science, whereas AI-enabled companies are implementation and distribution machines. The two company phenotypes establish moats at different layers — AI-first companies innovate just above hardware, whereas AI-enabled companies create enterprise value at the application-level.







We can no longer afford to conflate AI-first and AI-enabled companies.

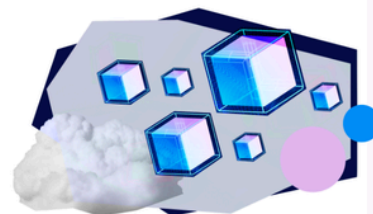
For founders, knowing what kind of company you are building is essential for recruiting proper talent, partnering with aligned investors, securing sufficient capital, and deploying a viable business model. AI-first companies require deep AI research acumen, investors willing to take a long view, materially more capital, and potentially less conventional business models than AI-enabled peers. In reality, this distinction is a spectrum, not a binary. Impactful companies will be built with both approaches, and far more will be AI-enabled than AI-first. For AI-first companies, though, we believe the fruit will be worth the labor. Influence over the technology stack from the ground up enables tight control over cost structure, immeasurable product optionality, and greater defensibility relative to AI-enabled companies that defer the exercise of scientific inquiry to those that are AI-first.

So far, the largest AI-first companies have been new entrants building for horizontal applications (e.g., [OpenAI](#), [Anthropic](#), [Perplexity](#)), countered by big tech product launches (e.g., [Google's Bard](#), [Amazon's Q](#)). Yet to be realized, vertical and industry-specific platforms, such as those in healthcare and life sciences, will showcase the expansive capabilities of large-scale models to deliver real-world impact. For founders, we believe enduring AI-first companies — in healthcare, life sciences, vertical industries, and beyond — will follow these six imperatives.

Six imperatives for AI-first companies

IN HEALTHCARE AND BEYOND

-  1. Create and sustain an undeniable data advantage
-  2. Recruit and empower AI scientists
-  3. Support a flexible AI stack
-  4. Establish distribution moats
-  5. Center safety and ethics in model development
-  6. Earn trust by solving real problems





1. Create and sustain an undeniable data advantage

AI-first companies exhibit an insatiable appetite for data and employ creative means for sustainable acquisition. However, more data is not always better, as researchers have observed diminishing returns in the context of AI scaling laws. Therefore, long-term differentiation in model performance requires multi-faceted approaches to data synthesis and curation, as well as optimized model architectures.

The data strategy must align with the model purpose.

For AI-first companies, the data strategy must align with the model purpose. The data required to support a generalist foundation model designed to be proficient in many tasks will differ from that powering smaller, task-specific models. The healthcare and life sciences industries may generate 30% of the world's data; however, the data exhaust of healthcare delivery alone is likely insufficient for developing highly performant systems. Intentional experimentation in data generation, coupled with elegant product design that facilitates model training with every user interaction, can bolster the quality of data for model training.

AI-first companies consider five key criteria to assess data defensibility:

- 1. Scalability: is it possible to amass an asset substantial enough for large-scale model training?
- 2. Continuity: can the dataset can be re-sampled over time?
- 3. Propriety: how easily can the data be accessed?
- 4. Fit: is the data relevant and suitable for a given model or task?
- 5. Diversity: does the data adequately reflect future real-world scenarios?

IMPERATIVE 1

Create and sustain an undeniable data advantage

	SCALABILITY	CONTINUITY	PROPRIETY	FIT	DIVERSITY
Publicly Available Data	Red	Yellow	Red	Yellow	Yellow
Customer-Generated Data	Green	Green	Yellow	Yellow	Yellow
Designer Data	Yellow	Green	Green	Green	Red
Synthetic Data	Green	Yellow	Yellow	Green	Yellow
Annotated Data	Yellow	Green	Yellow	Green	Red
"Feedback as Data"	Yellow	Green	Green	Green	Yellow



Below are key data sources for AI-first companies:

Publicly Available Data: Publicly available data is self-explanatory: datasets that are freely accessible and open for public use. This type of data is ideal for initial model training and benchmarking due to its wide accessibility and sometimes, diversity. Publicly available data provides a rich source of information for a broad range of applications. However, this data often demonstrates a high potential for bias and lacking task-specificity. Additionally, because these datasets are widely used and may have been included in the training set for large-scale foundation models, risk of contamination and overfitting are noteworthy, which can result in poor performance in real-world applications or on zero or few-shot tasks. One example is [MedQA](#), an open-source data of multiple choice questions collected from professional medical board exams. Though MedQA has been utilized as a benchmark for the performance of biomedical AI models, extrapolation of high performance on MedQA tasks into real-world settings has not been well-characterized to-date.

[OpenEvidence](#) is an AI-first company that powers physician-grade clinical decision support leveraging large-scale models to parse full-text biomedical sources for immediate answers to questions about drug dosing, side effects, curbside consults, treatment plans, and more. The platform enables efficient querying of publicly available biomedical literature via a natural language interface.

Customer-Generated Data: Customer-generated data is generated by clients of AI models during their normal course of business. This data is highly relevant to user needs and reflects real-world use cases, making it invaluable for training AI models in a dynamic and context-specific manner. Customer-generated data can also provide continuity, allowing for the development of up-to-date and responsive AI models. Nevertheless, customer-generated data can lack diversity as it can be biased towards certain customers (e.g., training on medical records from a community hospital in a rural vs. urban area), and might raise privacy concerns if synthetic techniques prove insufficient. Integral to the success of training on customer data is sampling from a sufficiently large and diversified set of customers in the curation process.

[Artisight](#), an AI-first company developing a suite of smart hospital products, utilizes synthetic derivatives of audio and video footage from hospital customers to train small and large models across computer vision and documentation-based tasks, addressing clinical workforce shortages and burnout.

Designer Data: Designer data sets are specifically created for model training use cases through intentional experimentation and are not typically found in publicly available or customer-generated datasets. These datasets are machine-readable and scalable and are designed to be highly specific and relevant, thus filling gaps in the aforementioned data sources. Designer data sets are particularly valuable for niche applications where standard datasets may fall short or where real-world settings do not generate data via a particular mechanism or in a specific form factor. Despite these benefits, creating designer data can be both expensive and time-consuming. If deployed in isolation, these datasets may lack generalizability.

[Subtle Medical](#), an AI-first company focused on imaging acceleration, generated millions of imperfect MRI images captured in 15 minutes, which were later utilized to train deep learning models that could reconstruct and de-noise medical imaging exams taken in shorter periods of time. In practice, imperfect MRI images provide little clinical value; however, for Subtle, these images trained deep neural networks that created a data moat for the company's technology.

Synthetic Data: Synthetic data is a subtype of designer data that is generated through simulations or statistical models that sample from other data sources, which can be highly useful in settings where real data is sparse or sensitive. Synthetic data also reduces privacy concerns. Challenges with synthetic data stem from concerns that it may not accurately capture the complexity of real-world data. Ensuring the accuracy and relevance of synthetic data is often a complex task, which can limit its effectiveness in training AI models.



[Unlearn.ai](#), an AI-first company focused on clinical trial acceleration, builds "digital twins," or synthetic clinical records that are capable of reflecting accurate, comprehensive forecasts of a patient's health over time under relevant scenarios.

Annotated Data: Annotated data sets are enriched with specific metadata or labels, providing additional context and detail. This enhancement of data quality and specificity makes annotated data particularly suited for supervised learning tasks, where detailed and accurate labels are crucial. However, the process of data annotation can be time-consuming and costly, and also carries the risk of human error in labeling, which can impact model performance negatively. In medicine, a field characterized as both an art and a science, "ground truth" labels often inject subjectivity, especially for "clinical diagnoses," where objective biomarkers do not yet exist. Promising research has also demonstrated the ability of AI systems to perform expert labeling tasks to mitigate these challenges.

[PathAI](#) and [Paige.ai](#), AI-first companies focused on pathology, leveraged human pathologists to annotate histopathology slides via a robust network of experts.

"Feedback as Data": Feedback as data, which refers to reinforcement learning with human feedback (RLHF), involves using human-user feedback as a direct input for training AI models. This approach allows models to adapt to complex and nuanced real-world scenarios and align more closely with human preferences and judgments. Many AI-first companies create user interfaces that enable model training via RLHF within the application as a byproduct of usage, agnostic to the underlying infrastructure. However, this method is subject to human biases and is limited in terms of scalability. The quality and effectiveness of the AI models are heavily dependent on the quality of the feedback provided, which can be variable and challenging to standardize. Model-based reinforcement learning, where models provide feedback to other models, is a flourishing area of contemporary research and may offer more scalable systems for fine-tuning.

[Abridge](#), an AI-first company that provides ambient documentation tools for clinicians, leverages clinician feedback on AI-authored notes to enhance note accuracy and quality across specialties.

Different data sets can and should serve distinct roles. Training data, used to teach the model fundamental patterns by providing a broad and varied base of information, will differ from fine-tuning data, which is more specialized, aimed at refining the model's performance in specific areas or tasks and addressing any deficiencies or biases the initial training may have overlooked. Evaluation data, separate from the training and fine-tuning sets, is crucial for testing the model's ability to generalize to new, unseen situations.

AI-first companies must balance data quality and quantity to optimize model performance. Multi-modal approaches that capitalize on the relative strengths and weaknesses of various data sources are likely to yield more enduring benefits than monolithic strategies alone.



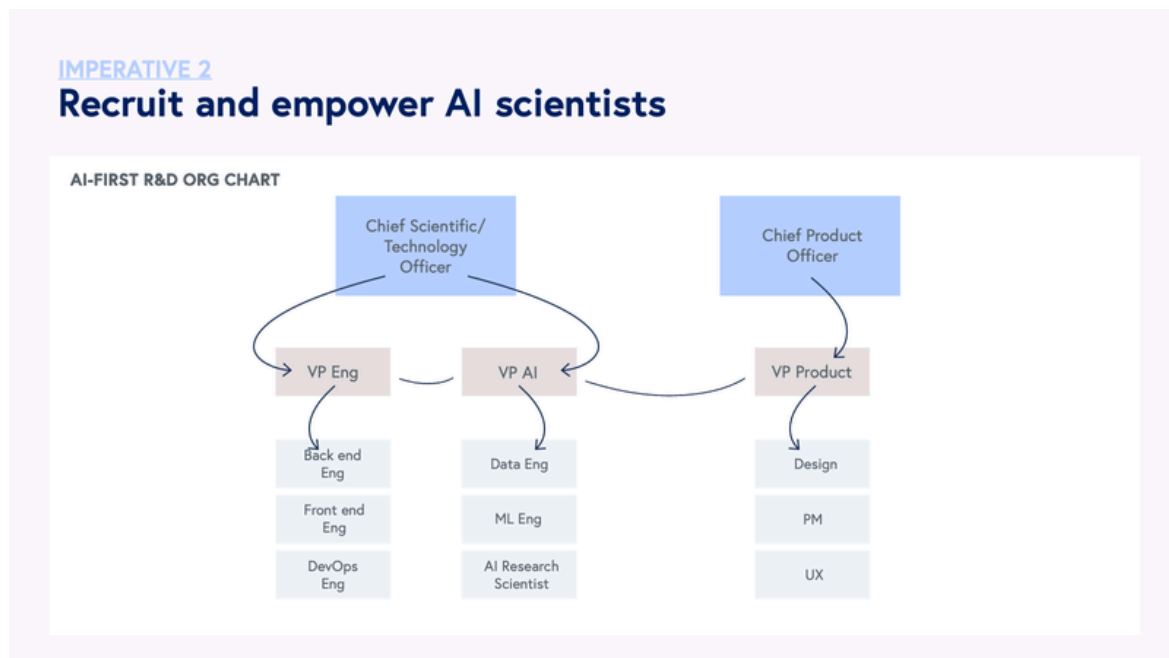
Data Type	Description	Strengths	Weaknesses	Example
Publicly Available	Datasets that are freely accessible and open for public use.	Widely accessible, diverse in nature, good for initial model training and benchmarking.	May contain biases, not always specific or detailed for particular needs, risk of overfitting.	NIH Clinical Trials database
Customer-generated	Data created by customers during interaction with a product or service.	Highly relevant to user needs, reflects real-world use cases, and delivers continuous data generation.	May lack diversity, potential privacy concerns, bias towards certain user groups.	Artisight
Designer	Custom-created datasets, tailored for specific AI tasks and not readily available in public.	Highly specific and relevant, can be created to fill gaps in existing data.	Expensive and time-consuming to create, may lack generalizability.	Subtle
Synthetic	Artificial data generated through simulations or statistical models, mimicking real data.	Useful in situations where real data is sparse or sensitive, no privacy concerns.	May not capture the complexity of real data, challenging to ensure accuracy and relevance.	Unlearn.ai
Annotation	Datasets enriched with specific metadata or labels, providing additional context.	Enhanced data quality and specificity, ideal for supervised learning tasks.	Time-consuming and costly to annotate, potential for human error in labeling.	PathAI, Paige.ai
Feedback as Data	Data derived from human feedback, used in reinforcement learning models.	Allows models to adapt to complex, real-world scenarios, and human preferences.	Subject to human bias, limited scalability, dependent on quality of feedback.	Abridge



2. Recruit and empower AI scientists

AI-first companies require "multilingual" teams — meaning they employ scientists deeply skilled in AI research as well as individuals with industry and business expertise. The design of the team must reflect advancement of AI as a key business activity. In healthcare and life sciences, this might take the form of clinicians and scientists partnering with AI researchers to design models with context-aware representations for a given domain. "Interpreters" are essential for the success of multilingual teams -- these are the rare individuals who boast interdisciplinary domain expertise, such as physician-informaticists, who can align various functional areas using shared rhetoric. For these reasons, AI-first companies are also more likely to benefit from an academic or industry laboratory affiliation. Atropos Health, a company focused on real world data generation for medicine, initially spun out of a Stanford AI Lab helmed by Dr. Nigam Shah.

The organizational structure must also reflect an AI-first company's prioritization of AI from the most senior levels. The R&D organization at an AI-first company will likely operate under a different reporting structure and leadership profile compared to an AI-enabled company. AI-first companies are more likely to have a Chief Scientific Officer (CSO) with deep AI research experience, with AI researchers and software engineering resources reporting into the CSO. AI-enabled companies are more likely to have Chief Technology Officers with classical software engineering training. Below is a sample organizational chart of how an AI-first company might structure their R&D team. For AI-first companies, we highlight the parallelization of a traditional software engineering org focused on application development and an AI research org focused on methods development and model fine-tuning.



The marketing function at AI-first companies also serves AI research as a key business activity. AI-first companies publish work regularly via accessible formats such as peer-reviewed journals (e.g., Nature Machine Intelligence or New England Journal of Medicine AI) or presentations at leading AI conferences (e.g., NeurIPS and ICML). These activities are critical for demonstrating advancements in state of the art (SOTA) and contributing broadly to advancements in the field of AI.



3. Support a flexible AI stack

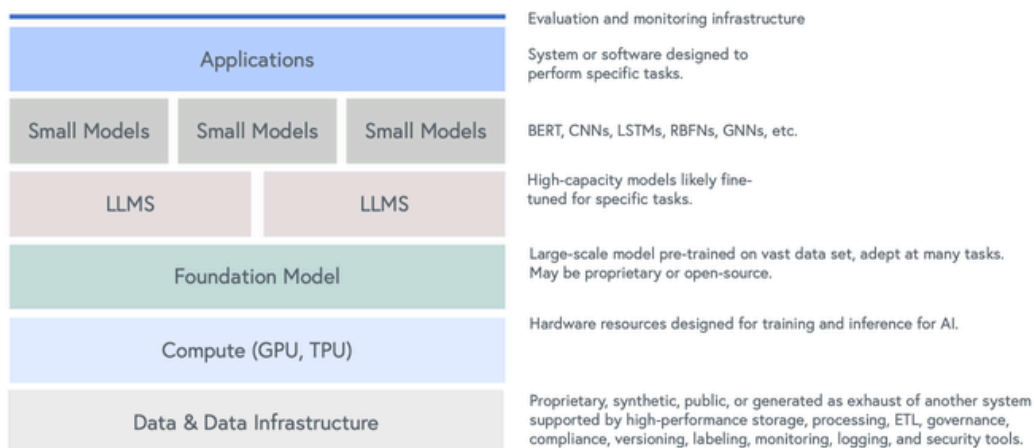
AI is advancing at an exponential pace as model sizes scale non-linearly and facilitate new and emergent behaviors that extend far beyond generation. By the time this blog post goes live, parts of it will be out of date. Critical for survival, AI-first companies must refrain from making rigid, irreversible, or monolithic decisions about the AI stack. Moreover, efforts to develop models that perform above SOTA benchmarks for every aspect of the stack is likely intractable for a single team. Instead, AI-first companies build modular AI stacks that leverage publicly available models (open-source, e.g. models shared on the Hugging Face Hub, and/or closed-source, e.g. GPT-4 or Claude 2.1) where others boast best-in-class performance, focusing internal proprietary model development resources in layers of the stack where the team has clear advantages due to undeniable data moats, methodological intellectual property, or other contributions of company or affiliated laboratory research.

We underscore the growing importance of open-source models across industries, but especially in healthcare and life sciences, where after many years of closed-system architectures, transparency is emerging as a core value for technological innovation. Open-source approaches grant full control over all aspects of a model and its training data, including forward-looking updates. AI-enabled companies are more likely to rely on incumbent infrastructure, such as GPT, for a majority of AI-related product features. These companies may perform fine-tuning on top of these models, but as a result, are more likely to hit a ceiling in terms of product capabilities. By innovating lower in the AI stack, AI-first companies enjoy greater product and feature optionality over time.

Furthermore, not all AI-first companies are building foundation models. Instead, many focus on developing smaller specialized models, which are characterized by fewer parameters. These models have been shown to challenge scaling laws for large-scale models by 1) outperforming large models on particular tasks, and 2) demonstrating superior performance when deployed in combination with larger models. By supporting a flexible stack, AI-first companies can enjoy the strengths of large models in orchestrating utilization of small models -- managing, coordinating, or directing fleets of small models to address particular tasks -- akin to an "AI project manager." From a business perspective, small models offer other benefits such as reduced costs, lower latency, and greater controllability.

IMPERATIVE 3

Support a flexible AI stack





4. Establish distribution moats

Without distribution, the impact of a best-in-class AI model may start and end with a published paper. Both AI-first and AI-enabled companies should seek to obtain distribution moats or advantages early on, such that it's conceivable the product will be able to integrate with or displace incumbent technology providers over the near-term. Unlike AI-first companies, however, distribution is often the only moat for AI-enabled companies. AI-first companies enjoy both technical and distribution moats, both of which contribute to enterprise value creation.

Though direct sales can be a critical part of go-to-market (GTM), accelerating uptake of AI-first products in the enterprise often requires more creative strategies:

1. Product-led growth: leveraging the product for user acquisition and retention.

OpenEvidence, a biomedical AI company, partnered with Elsevier, a leading medical publisher, to launch ClinicalKey AI, a clinical decision support tool that will deploy Elsevier's trusted, evidence-based medical information via a natural language interface.

2. Partnerships: collaborations between organizations for mutual benefits such as preferred vendor relationships.

Abridge, an AI-first company focused on tools for the clinical workforce, partnered with Epic, the leading electronic health record provider, as a preferred vendor partner for ambient documentation solutions.

3. White-labeling: rebranding and selling AI-first technology within a partner company's product or service.

Iterative Health, an AI-first precision gastroenterology pioneering novel biomarkers for gastrointestinal disease, secured an exclusive partnership with Provation, the leading gastroenterology electronic medical record, to facilitate broad distribution in the specialty.

IMPERATIVE 4

Establish distribution moats

DIRECT SALES

Enterprise sales team establishes category-leadership among champions and budget-owners.



PRODUCT-LED GROWTH

High quality product and favorable user experience drive customer acquisition, adoption, and retention.



PARTNERSHIP

Synergistic relationships with incumbents that unlock access to a primed customer base.



WHITE-LABEL

Offering AI products or services that can be rebranded and resold by other businesses in an embedded format.





Distribution advantages are key, but so is the answer to the question: who pays? Payment models remain an outstanding question for AI-first companies, especially in the biomedical realm. Many emerging companies are leveraging conventional business models such as software-as-a-service and transaction-based revenue models, yet healthcare stakeholders cite cost as the top barrier to AI implementation. The healthcare industry does not have adequate resources or aligned incentive structures to subsidize a transition to AI. By some estimates, hospitals only spend \$36 billion annually on all IT investments spanning software, hardware, and human resources. Payment reform will be critical.

In the interim, we expect commercial success for AI companies that deliver on Law 5 of Bessemer's 10 Laws of Healthcare: demonstrating financial and clinical ROI. Despite our excitement for emerging methods that allow for improved diagnostics, we fear these applications will be slow to garner adoption unless the improvement in clinical outcomes drives a corresponding financial benefit, as would be the case in a value-based organization, but potentially less so in a traditional fee-for-service provider.

5. Center safety and ethics in model development

As AI permeates all aspects of public and private life, AI-first and AI-enabled companies must grapple with preserving the foundational rights of human users. Investment in AI safety and ethics is non-negotiable for AI-first companies innovating at the foundational layer. These companies must exercise caution and intention in data custodianship and a relentless commitment to model maintenance. Strategies such as continuous performance monitoring, fail-safes and overrides for human intervention, recurring re-validation against real world data, and user training that outlines key limitations of AI are employed exhaustively by AI-first companies.

In conjunction with the Coalition for Health AI (CHAI), we recognize that safety and ethics begins at the earliest stages of model development design all the way through monitoring and real world application.

At the heart of CHAI's mission is the development of comprehensive guidelines and guardrails for healthcare AI technologies. With a principle-based approach, CHAI is spearheading efforts to harmonize existing standards and, in collaboration with the healthcare AI community and validation labs anointed by the Biden Administration's Executive Order, establish common principles that will serve as the bedrock for creating and implementing trustworthy AI practices.

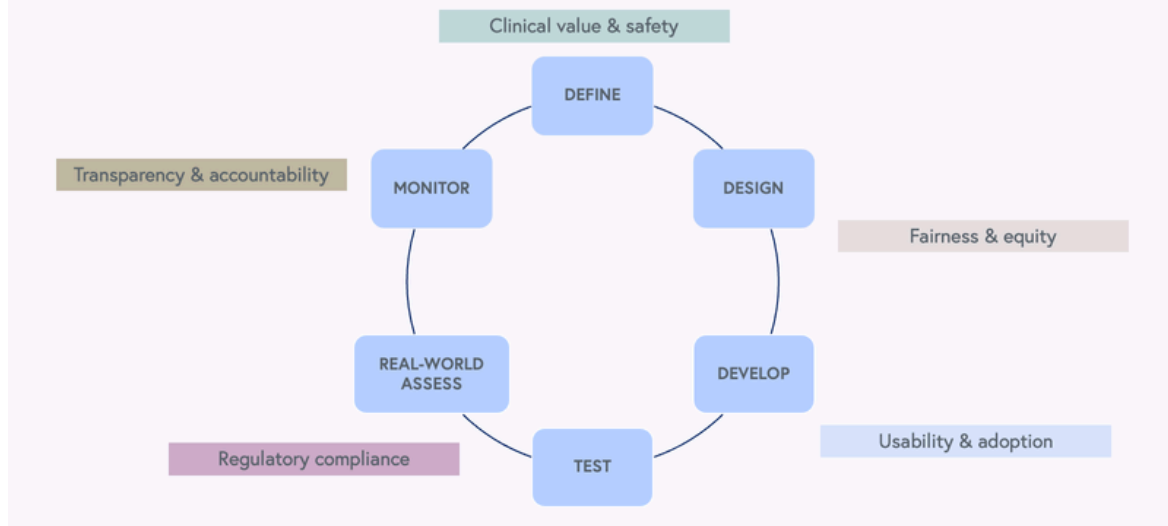
These values include:

- Safety: AI systems must not put human life, health, property, or the environment at risk.
- Accountability and transparency: Individuals involved in the development, deployment, and maintenance of AI systems must maintain auditability, minimize harm, report negative impacts, and communicate design tradeoffs.
- Fairness with bias management: AI systems should manage bias effectively to ensure disparate performance or outcomes for selected groups are minimized.
- Security and resilience: AI systems should be able to withstand adverse events or changes and maintain their functions and structure.
- Privacy enhancement: AI systems should adhere to norms and practices that safeguard human autonomy, identity, and dignity, in compliance with relevant privacy laws and standards.



IMPERATIVE 5

Center safety and ethics in model development



6. Earn trust by solving real problems

While AI-first companies are built on data, they thrive on trust. Trust is earned when stakeholders perceive that their problems are met with curiosity and empathy rather than techno-solutionism. It is earned through reliability, accuracy, and respect for the human condition and a focus on value creation for society. AI is one of many technologies we have for solving societal problems. There is a difference between what AI is capable of and where AI creates value.

Whether it makes sense to leverage AI is specific to the problem, task, and industry. AI is both a means to an end – many patients do not care whether a drug was discovered by AI or by human scientific intuition, they care that there is an accessible therapy with a favorable side effect profile that may treat their condition – and also a tool for solutioning – re-designing clinician user experience leveraging ambient data capture to address administrative burden and burnout.

As with any emerging technology, we must earn the right to use it, and we ought to demonstrate superiority with the way things have been done before. AI-first companies have a tall-order – they must innovate technologically and demonstrate an ability to solve real-world problems, all while operating in alignment with the value systems that govern society.

Setting the record straight

Building AI-first companies, especially in healthcare and life sciences, is not an easy feat. However, the impact of AI-first companies will be greater, financial returns superior, and moats more enduring than their AI-enabled counterparts. Though we're in the earliest days of witnessing AI-first companies in the wild, including industry-specific opportunities in healthcare and life sciences, the stark contrast in the approach, capabilities, and leadership of AI-first companies clearly distinguishes them from traditional software or AI-enabled businesses. AI research will continue to be the lifeblood of generational opportunities in AI. It is imperative that the venture capital and startup ecosystems commit to distinguishing AI-first companies from AI-enabled and become students of how these companies are built and scaled. As we continue to interrogate the role of AI as an agent for problem-solving, we implore founders to think deeply about what kind of AI company they seek to build and what that will mean for the path ahead. Perhaps most importantly, let us not mistake a clear view for a short distance. This is just the beginning.

Seven product strategies to prevent churn for B2B AI app leaders

Janelle Teng and Sameer Dholakia



Venture insights that matter

bvp.com/subscribe

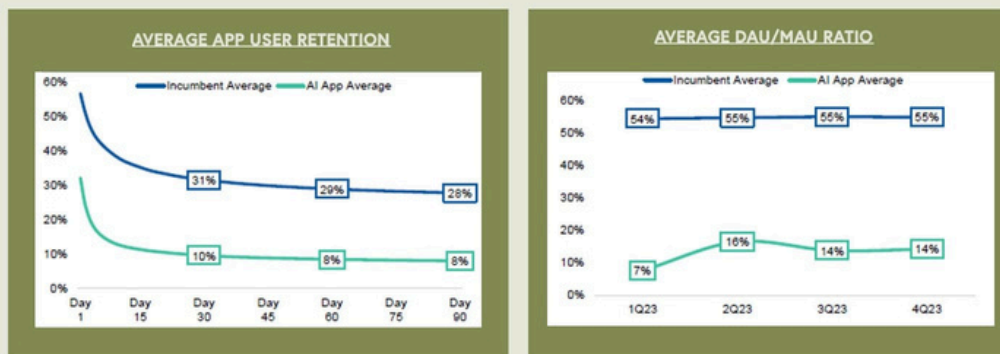


Preemptive ways AI app builders can safeguard their business from the pernicious leaky bucket syndrome.

The release of ChatGPT in late 2022 unleashed the floodgates on mainstream AI adoption by putting the power of large language models directly in the hands of everyday consumers. Over [100 million people had experienced AI](#) within mere months of the launch, and businesses soon followed suit. The tech industry saw a surge of AI-native and AI-embedded consumer applications, prompting [OpenAI to launch the GPT store](#) earlier this year. Many consumers around the world have now tried ChatGPT or other B2C AI apps — but did these users stick around?

With over a year of data on consumer behavior related to B2C AI apps, an early trend has emerged — while consumer AI apps have experienced strong top-of-funnel growth and accelerated user acquisition, these businesses have also experienced a notable "leaky bucket" problem. User retention and engagement metrics of consumer AI apps are markedly weaker than incumbent consumer apps:

Consumer AI Apps show weaker retention and engagement compared to incumbent consumer apps



Source: Morgan Stanley Report (December 2023)

In other words, churn is one of the most prevalent issues plaguing consumer AI businesses as they struggle to retain long-term customers and drive habitual usage. There are several plausible reasons for this phenomenon, including a growing competitive landscape as the democratization of AI model APIs have lowered barriers to entry for B2C AI startups.

While every new market opportunity comes with initial growing pains and adjustment periods, churn is not a concern to be taken lightly. Eventually AI novelty will wane and the growth momentum of these types of businesses will slow down. It will be significantly harder to maintain growth if B2C AI companies are struggling to plug a sizable churn gap every month.

As enterprise investors, we often observe how B2C trends could map to B2B businesses. So a key question emerges: Will B2B AI Apps suffer from the same churn fate seen in B2C AI Apps?



Enterprise AI adoption has lagged consumer AI adoption, which is not surprising given the higher enterprise requirements around governance, security, integrations, and output quality. Consequently, there tends to be a slower ramp for enterprises in adopting new technology – many corporations only started experimenting with AI applications through proof-of-concepts or pilots within the past few months, and are not expecting full-scale rollouts till later this year.

Since enterprise AI adoption is in its early innings and many B2B AI companies have yet to hit their first renewal anniversaries, there has not necessarily been sufficient longitudinal and cohort data to credibly determine B2B AI retention benchmarks. So the leaky bucket phenomenon we see in B2C has not become apparent in B2B yet, but that risk remains.

One source of relief is that B2B software companies generally demonstrate strong net dollar retention dynamics, as we posit in [how to scale to \\$100 million](#). But given the precedent set by consumer AI companies, we still don't know if these incumbent B2B retention benchmarks will hold for B2B AI companies. And consumer AI's churn issues have certainly raised the guard for enterprise AI companies, as some early warning signs are already emerging amongst B2B AI app companies that are mainly perceived to be "wrappers" on top of GPT or other models. This retention concern is compounded for B2B AI companies that primarily leverage a prosumer model or rely on self-serve to drive enterprise top-of-funnel, since those models can sometimes share similar characteristics to consumer business models.

Acting swiftly upon these signals, we've seen a bold cohort of B2B AI app first-movers lead the charge on leveraging product strategy as a pre-emptive measure to boost defensibility.

There are certainly many approaches to build moats, such as through pricing levers or go-to-market strategy. But as former product people, we're advocates for thoughtful product strategy as a powerful way for B2B AI apps to improve user engagement and retention. And we're not simply referring to product vectors that are table-stakes such as effective UI/UX or strong product velocity. Rather, we've observed several B2B AI app leaders, both [AI-native](#) as well as [embedded AI app companies](#), craft extremely deliberate roadmap strategies to enable their AI products to avoid the leaky bucket pitfall.

Here are seven lessons with case studies from inside and outside of Bessemer's portfolio that demonstrate these best practices for B2B AI app businesses:

Seven product strategies to prevent churn for B2B AI app leaders

1. Embed into incumbent platforms through integrations and partnerships
2. Meet users where they are
3. Generate a tangible work product
4. Build across the chain to deliver more value to your ICPs
5. Leverage proprietary data or novel data techniques to create moats
6. Go multi-modal
7. Maximize network effects through platform architecture

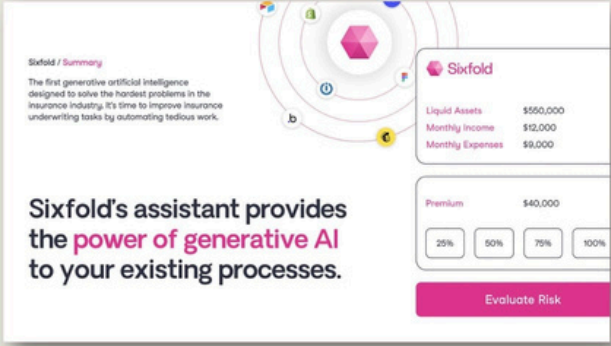


1. Embed into incumbent platforms through integrations and partnerships

B2B AI app startups have to start somewhere. It's common to see companies go-to-market initially with a point solution, aiming to grow into a platform as the company matures. That said, as a point solution, it can sometimes be challenging to become deeply embedded into existing workflows. This poses a risk to user retention and engagement. However, we've seen B2B AI startups get resourceful here by partnering with other platforms, especially large incumbents, for leverage through tight integrations.

1. Embed into incumbent platforms through integrations and partnerships

CASE
Sixfold



Sixfold / Summary
The first generative artificial intelligence designed to solve the hardest problems in the insurance industry. It's time to improve insurance underwriting tasks by automating tedious work.

Sixfold's assistant provides the **power of generative AI** to your existing processes.

Liquid Assets	\$550,000
Monthly Income	\$12,000
Monthly Expenses	\$8,000

Premium: \$40,000

25% 50% 75% 100%

Evaluate Risk

Sixfold is cleverly designed to be embedded as an API or plug-in into an underwriter's existing PAS, allowing its co-pilot to be directly integrated into an underwriter's daily workflows.

Bessemer portfolio company [Sixfold](#) – a leader in generative AI for insurance underwriters designed to boost underwriting capacity, accuracy, and transparency -- is a role model on this front. Underwriters typically work in Policy Administration Systems (PAS), which help them rate, quote, and bind insurance policies with insurance agents. Sixfold is cleverly designed to be embedded as an API or plug-in into an underwriter's existing PAS, such that insurers do not need to overhaul legacy systems or replatform their workbench in order to interact with Sixfold's co-pilot. Consequently, underwriters can experience the power of Sixfold's AI very effortlessly since Sixfold is seamlessly integrated into existing daily workflows.

2. Meet users where they are


Related to embedding into workflows, B2B AI applications drive higher retention when their product meets users where they are. Unsurprisingly, usage and engagement metrics tend to drop off whenever users are faced with unnecessary friction to access a product (e.g., multiple sign-in portals, complicated click paths, or additional windows to open) as time-to-value becomes longer. To mitigate this, many B2B AI app companies make their products directly accessible within platforms that their customers use frequently as part of a daily workflow.



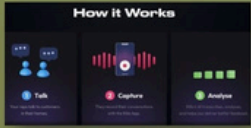
This could look different given your customer base and the technology you're providing, but here are four exemplary case studies:

2. Meet users where they are

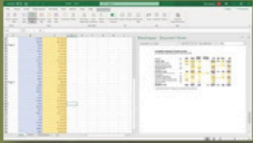
CASE
Databook
Personal AI sales advisor can be accessed using organizational communications tools like Slack and Teams so that reps do not need to toggle between platforms



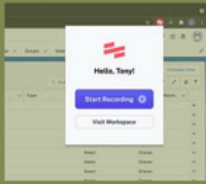
CASE
RILLA
Offers a mobile app for AI powered ridealongs, purpose-built for field sales teams who are often out-and-about and not desk-bound



CASE
datasnipper
Audit and finance professionals can leverage features directly within Excel through a plug-in



CASE
Scribe
Provides a browser extension so that users can create and view step-by-step guides, automatically and in-context, in any application as they work



Source: Company websites

- Bessemer portfolio company Databook's personal AI sales advisor can be accessed using organizational communications tools like Slack and Microsoft's Copilot for Sales so that reps do not need to toggle between platforms.
- Rilla, a leader in virtual ridealongs for outside sales, offers a mobile app for AI powered ridealongs, purpose-built for field sales teams who are often out-and-about and not desk-bound.
- Datasnipper's Intelligent Automation Platform can be accessed directly by audit and finance professionals within Excel through a plug-in.
- Scribe provides a browser extension so that users can automatically create and view step-by-step guides, in-context in any application as they work. Scribe's Sidekick lives on a user's panel and directly surfaces Scribe content relevant to whatever page or application they are working on.

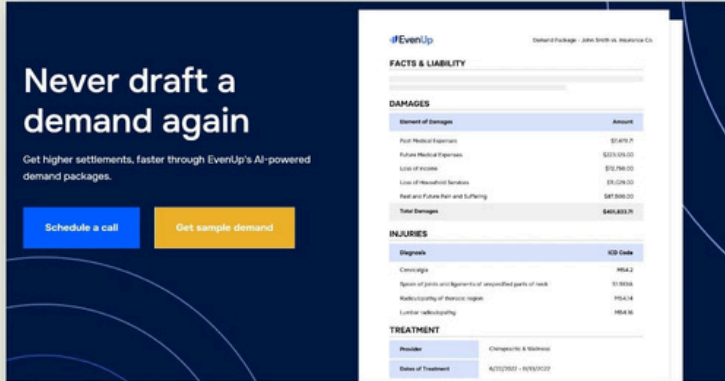
3. Generate a tangible work product offering

Enterprise workflows often involve the creation of a tangible work product. This could be a piece of analysis, document, or a report. One very effective way for B2B AI apps to lock in stickiness is to leverage AI to generate this outcome of a core work process.

Bessemer portfolio company EvenUp, which provides Demand Packages for Personal Injury Lawyers, exemplifies this strategy. Demand letters are a core yet time-intensive component of filing personal injury claims. EvenUp leverages AI, coupled with their proprietary database of settlement data, to automate the creation of demand letters. With this capability, EvenUp helps to free up precious time so lawyers can take on more cases and spend more time with their clients, instead of manually drafting letters. Additionally, because EvenUp's one-of-a-kind database can locate relevant high-dollar-settlement precedents, its AI-generated letters contain unique value-add data points that help drive more revenue for attorneys.



3. Generate a tangible work product



EvenUp leverages AI and their proprietary database of settlement data to auto-generate high-quality demand letters.

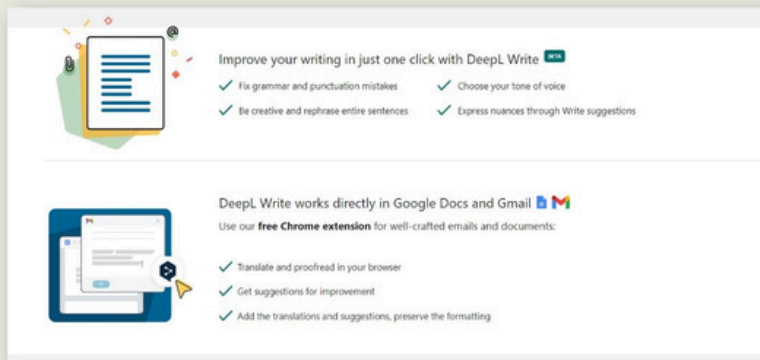
Source: Company website

4. Build across the chain to deliver more value to your ICPs

After conquering a specific part of the workflow, we've seen B2B AI apps increase defensibility by moving across the value chain horizontally in order to deepen the relationship with users of an ideal customer profile (ICP).

Bessemer portfolio company DeepL, a full-stack neural machine translation service, exemplifies this principle. The company's first offering involved large-transformer based NLP capabilities, yielding translation improvements across languages. Since then, DeepL has expanded its offering suite to adjacent capabilities beyond translation, including DeepL Write which was launched early last year as an AI-writing companion.

4. Build across the chain to deliver more value to your ICPs



DeepL's first offering involved large-transformer based NLP capabilities. Since then, the company has expanded its offering suite to adjacent capabilities beyond translation, including writing.

Source: Company website



DeepL is a role model for how to evolve into a full-stack platform by building for a user's broader value chain, ultimately increasing a business' defensibility. Essentially what this means is that with each feature extension, a business reinforces value or extends product capabilities for its original ICPs.

5. Leverage proprietary data or novel data techniques to create moats

As B2B AI App leaders move across the stack, they don't just reinforce ownership of the relationship with the end user, but can begin to shift upstream to reach a highly strategic part of the stack: the data layer.

This positioning can help to drive defensibility. For example, EvenUp has an intimate understanding of how different health systems and hospitals create their medical bills, so their AI can continuously get better at accurately parsing their billing information to feed their model to draft demand letters.

In another example, Databook augments public data (e.g., 10-Ks, annual reports, and earnings transcripts) and customer data (e.g., opportunities and use cases) with proprietary data including contact data, strategic priorities, and technographics. To enhance these data moats, they then apply proprietary computational analysis that feed into LLMs to deliver powerful account intelligence insights to customers.

Shift Technology is also a strong role model for how to leverage unique data techniques to build competitive advantage. Shift automates and optimizes critical insurance decisions with best-in-class AI powered by a novel unified data approach. Shift recognizes that on a first principles basis, better data can inform better decisions. Thus, their platform takes on the resource-intensive work of mapping insurer data, such as policy and claims data, with the best external data sources, such as government records and publicly available social media, resulting in a single, powerful unified data set for risk detection and claims automation.

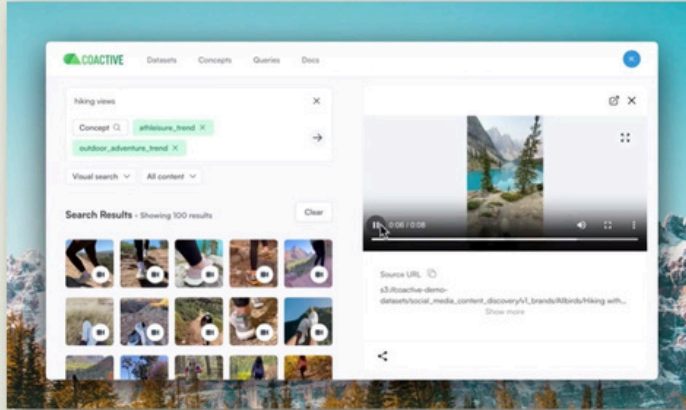
6. Go multi-modal

Today, enterprises face a proliferation of data sources and formats, so as B2B AI App leaders think about product expansion, having a multi-modal offering is often another effective way to extend platform capabilities into adjacent areas.

A great case study here is from Bessemer portfolio company Jasper. A primary early use case for marketers adopting Jasper's AI solution was the creation of long-form blog posts using text-based GenAI capabilities. Naturally, once the post was generated by AI and edited by a marketer, the next logical step in the workflow was to find appropriate imagery to bring the blog post to life. So Jasper recently completed the acquisition of Clickdrop to strengthen its Jasper Art product, using multi-modal capabilities (both text and image) to address a marketer's entire value chain.



6. Go multi-modal



Source: Company website

Going multimodal across all visual mediums, including image and video, enables Coactive to deliver more value to customers by enabling users to unleash the full potential of their entire visual content library within a single platform.

Another role model here is Bessemer portfolio company [Coactive.ai](https://coactive.ai) that helps enterprises to derive key business insights from all types of visual content, including video and images, using a data-centric approach. Visual data has traditionally been an under-tapped resource at companies since incumbent tooling is often limited in capabilities to handle unstructured data. Going multimodal across all visual mediums enables Coactive to deliver more value to customers by enabling users to unleash the full potential of these content sources within a single platform. For instance, customers can run AI-powered search or create concepts across their entire visual content library from Coactive's platform in order to gain a holistic view across different use cases.

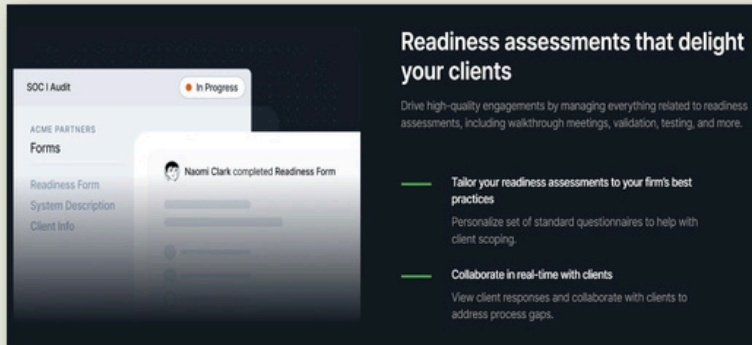
7. Maximize network effects through platform architecture

Among developer SaaS platforms and PLG leaders, we've seen how creating a flywheel of network effects is a powerful way to drive stickiness — whether it's in a pricing strategy, go-to-market approach, or features within the product. This perennial principle on the power of network effects holds true for emergent B2B AI apps as well. Bessemer portfolio company, [Fieldguide](https://fieldguide.ai), (AI built for advisory and audit) illustrates how to execute on this tenet, as its platform architecture is strategically built with both practitioner and client-facing interfaces in order to drive and maximize network effects.



7. Maximize network effects through platform architecture

CASE
FIELDGUIDE



Fieldguide's platform is strategically built with both a practitioner and client facing interface in order to drive and maximize network effects.

Source: Company website

Fieldguide simplifies engagement management by replacing a multitude of point solutions with one end-to-end, AI-powered platform. This provides practitioners with all of the functionality needed across the engagement lifecycle in one place - from testing and controls to client requests, report writing, and analytics. Most importantly, integrated collaboration tools enable teams to coordinate and share documents easily, fostering networks within organizations.

Moreover, practitioners can invite their end-user clients, such as CEOs, CFOs, risk and compliance leaders, CISOs, and Internal Audit leaders, to access the Fieldguide platform through a secure, role-based client portal. This allows clients to respond to evidence requests, collaborate on documents, and stay updated on engagement progress in a streamlined and intuitive manner. Clients can conveniently view assigned information like requests, tasks, and comments without searching through emails or making manual edits. Bringing clients onto the platform enhances distribution virality.

Fieldguide exemplifies the power of a product that enables both sides of the market to actively engage with a single platform. This dual engagement fuels sharing, boosts engagement levels, drives adoption rates, and ultimately amplifies the value for all participants within the network as time progresses.

User-centricity is key for app layer companies

These case studies illustrate how product strategy can play a critical role in mitigating churn for B2B AI applications. A unifying theme amongst all these examples is that B2B AI app companies must build products with a deep level of user empathy and awareness, as every industry or role has specific workflows and nuances.

Ultimately, application layer companies are in such a unique position given that they sit closest to the end user. So we implore all B2B AI app companies to capitalize on this unique position and take full ownership of the user relationship to build out highly-beloved AI products!

How Intercom navigated the AI paradigm shift



Venture insights that matter

bvp.com/subscribe



Intercom's Des Traynor and Fergal Reid share field notes on what other builder teams should know as they transform their products with language models.

GPT-3 sparked the flame, and then GPT-4 set the software world on fire. For SaaS builders, capabilities enabled by the recent advancements in large language models mean that anywhere there is natural language processing (NLP)—within the application, within the product workflow, and within the ecosystem—there is an opportunity to leverage these massive predictive calculators to reimagine SaaS solutions.

"The first and most important thing for SaaS leaders to understand is what is now possible that wasn't before," Des Traynor, co-founder and chief strategy officer of Intercom said. "With GPT-4, for example, it is now possible to generate text, generate images, infer meaning, expand, contract, and take actions. So how can builders make software products easier, faster, and more productive, across different use cases?"

Long before ChatGPT became a household name, Intercom, a leading customer service platform, had been building Artificial Intelligence (AI) into their product since 2018 and experimenting with how it could transform customer support. As of today, the Intercom team has launched two products using OpenAI's technology, [Inbox AI features](#), which among other things allows support reps to compose, summarize, and expand entire conversations with customers with the click of a button, and [Fin](#), their GPT-4 powered customer support bot.

Amidst this advanced computing era, the tech ecosystem is exploring applications and asking tactical questions on how exactly to reimagine products and drive value with LLMs.

Des Traynor, co-founder and chief strategy officer of Intercom, and Fergal Reid, senior director of machine learning, share their field notes for other product and machine learning (ML) teams. Together, they distill five lessons they learned while building AI products that other SaaS leaders can apply to their future product roadmaps.

Pick a low-risk, high-potential problem

"Selection is the hardest challenge," said Des Traynor, co-founder and chief strategy officer of Intercom. "Finding an area that is relatively exploitable by generative AI, but also provides room for imperfection, is the challenge. That's why our first foray into generative AI was on the agent assist side (Smart Replies)."

In this case, Intercom was simply augmenting a workflow by suggesting answers to support agents, whether it was to condense, summarize, expand, or even change the tone of the answers they were already writing. "If it didn't work, there was zero downside. If it did work, there was significant upside," Des shared.

"No matter the use case, builders should initially focus on an asymmetric opportunity. And whatever you select, remember that there needs to be a somewhat regular and active use case. With low frequency, you'll never get a sense if it's working or not."

However, with this advice, Des warns that builders must assess and forecast downstream costs before building. A worst case scenario would look like adding expensive generative AI to an extremely high frequency and reasonably low-value feature. With millions of users, this experiment could bankrupt a business without driving real customer value.



Start with a small Skunk Works team to gain traction

Though Intercom had a long held belief that the future of customer support would one day be automation working alongside humans, there was still a lot to assess before building. The team had to define clear goals, identify use cases, and prioritize features.

Initially, Intercom kept this early team tiny so they could stay as agile as possible; in other words they gathered a Skunk Works team—a group of dedicated people interdisciplinary in their thinking and collaboration.

"Driven mostly by the ML team, we deliberately ran a very small operation with some product and beta support from outside," said Fergal Reid, leader in machine learning who has been with Intercom for over seven years. "This reduced overhead of internal communications in such tight timeframes. "As the technology was so new, we didn't want to waste any precious time with a large group ramping up and trying to get on the same page."

Truthfully, there were still plenty of internal obstacles around trying to run such a high velocity project. "However, with Des, we had great executive stakeholder support to help reduce the friction when it came to getting legal in place, beta terms and conditions, etc," said Fergal. Now in full production, and two launches later, there are many more teams surrounding AI launches, but early success started with a tight-knit team.

Figure out a ship-fast-and-learn framework

"For context, we've been building (non-generative) AI products for years," said Fergal. "Our end-user facing bot, Resolution Bot, has been resolving support queries automatically since 2020. Our team used neural networks to detect topics, so we were already utilizing AI features in the Inbox to make support reps more efficient."

This was just one early example—and as the efforts of their Skunk Works team continued, they had to create a framework for prioritization.

"Intercom has a culture around [shipping fast and learning](#), which is a major asset in times of technological disruption. Initially, we set out to build a series of features that required either thin or medium depth integrations, so we could deploy them quickly, but that would still deliver real value," Fergal explained.

The team thought language models would only extend to features that kept a human support rep in the loop. So, they started working on the "Inbox" features, building several generative features to help support reps write text in the Inbox, as well as to summarize conversations before handover. "If anything, we underestimated the value of these AI features."

"If anything, we underestimated the value that our customers would get from these early features." "As mentioned, our first feature of this type, [Smart Replies](#), shipped in June 2022. The Smart Reply feature makes it quicker and easier for a support rep to reach out to customers by offering predictive text—based on common greetings used in the past—which the customer can accept, reject, or alter for personalized engagement at speed," said Des.

The team was in the middle of building a next-gen version of the Smart Replies model, with the intention that it would run on in-house generative models, when ChatGPT launched.

"ChatGPT, and text-davinci-003, completely blew us away in terms of quality, so we instantly started to focus on these models, rather than our internal models," Fergal shared.



The team knew this was a moment to pivot. Swiftly they moved production from their proprietary models into using OpenAI's technology, leveraging GPT-3.5, GPT-4 and Plugins for better and faster outcomes.

Now that the team could outsource its model to GPT-4, their development could kick into high gear. "We could now use this foundational model to build a bot that talks to end users." This was the start of their latest product Fin, one of the first GPT-powered customer support bots on the market.

Centralize a machine learning team

"We are four to five months into a hurricane of opportunity, and it is very early days," said Des, "but I do think every software company will need significant expertise in AI and ML to identify the best opportunities specific to every business and to know how to explore and exploit."

Translation: If you've yet to map out an ML team, the time is nigh.

"I don't think it's realistic for every engineer in a tech company to overnight become an AI and ML expert. They will probably all become very familiar with the various APIs that are out there, but I think the research and exploration around opportunities is where there is the most strategic opportunity."

Intercom made it a priority to ensure the best practices of AI implementation extended across product, technology, and strategic teams and executive leaders alike.

Here's how Intercom breaks down the ML vs. Product and Engineering teams:

Team	Role	Focus
Centralized Machine Learning	<ul style="list-style-type: none">— Explore, exploit, present opportunities— Partner with teams to suggest ways in which AI could influence the roadmap or significantly rewrite certain features or change long-held assumptions about the world.	Be aware of the changing capabilities, challenge assumptions, and identify ways to effectively distribute capabilities and learnings across the organization.
Product and Engineering	<ul style="list-style-type: none">— Execute based on opportunities identified by ML teams— Build prototypes and new solutions that go into production	Find ways to build a new reality; tasked with taking a feature they previously deemed non-automatable and automating it.



The framework lightly outlines the breakdown between the different teams, but that doesn't mean it won't change in the future.

Still, there are must-have qualities that all emerging ML teams must possess: Curiosity, the ability to apply research to different contexts, and to think like a customer. "A ML team must be very deep in the culture of R&D, meaning they are used to working with new ambiguous technology, but also care deeply about solving customer problems, and moving and shipping extremely fast. That combination is powerful," said Des. "In 2022, we were already deeply working on the relevant customer problems, trying to use other technologies, and so when GPT-3.5 came along, we were really well positioned to move rapidly."

And according to Fergal, even the sharpest of ML teams will only be as successful as a company's long-term vision: "At Intercom, we see a future where AI will drive the arc of our technology progress—and this is the accelerant of our productivity. Because all teams were aligned with this company vision, it created a great deal of internal excitement and momentum."

Reimagine language as the new UI for software

Transforming your product suite with AI requires first-principles thinking and assessing the repetitive, predictable, and expandable tasks and workflows your business facilitates. "Once you understand the scope of your product you can see the extent to which AI can change your product," explains Des.

"Enterprise software has a history of clunky UIs; one massive question we'll be answering in the industry over the next two years is: Is text the new UI? And is voice the new UI?"

"During the little chatbot renaissance we saw in ~2015, people would've laughed at the notion that chat could be the next interface, but OpenAI, GPT-3.5+, and whisper, have allowed natural language to evolve human-computer interaction."

As a simple example, Des poses the power of text-driven UI for a tool like Google Analytics. "It would make it simpler and easier for new users to get productive within Google Analytics immediately because they don't need to learn all the dropdowns, settings, filters and segments, etc. And that would be extremely disruptive to enterprise web analytics because now everyone can use the product. With language, all you need to do is ask the right questions. Very shortly, and within the year (2023), we will see a lot of complex tooling that offer plain English UI (or any language) across not just written text, but also synthetic voice, real-time audio translation, and transcription."

"Voice-driven UI is next and will have long-lasting transformation for how and where software is considered." As Des explains, voice works in places where access to screens limited.

"We'll see the resurgence of the 'place-on-a' vs. the persona, meaning that UI will be truly multi-modal and anyone can engage or direct software through text, voice, and other mediums." What Des is describing for SaaS builders is a future where software can break from devices and become more relational and conversational.

"We're going to continue to use a complex sales dashboard or a CRM or product analytics, but instead now, you can do this work aloud while you're doing the dishes, getting out of your driverless car, and in a whole new set of contexts throughout our lives."

AI escape velocity: A conversation with Ray Kurzweil



Venture insights that matter

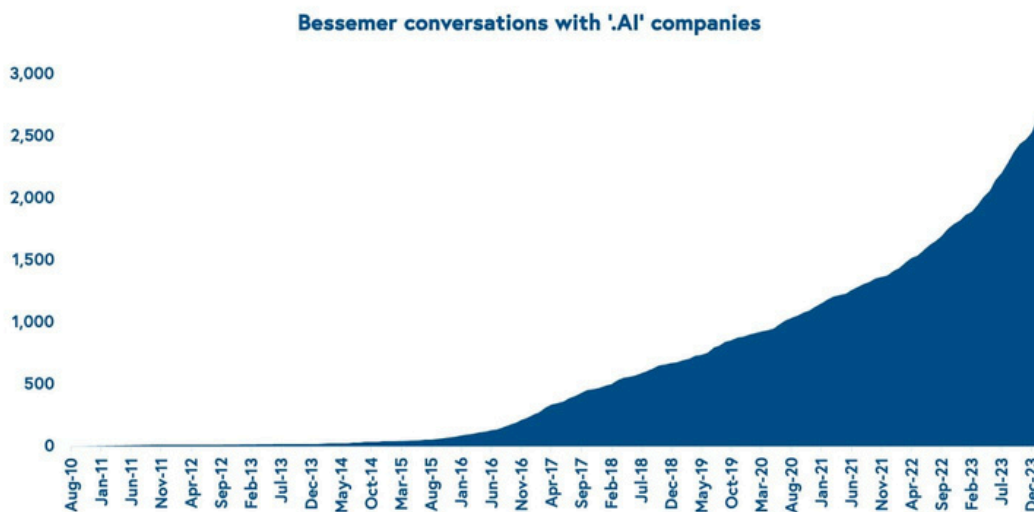
bvp.com/subscribe



Legendary Futurist Dr. Ray Kurzweil joins Bessemer to discuss why the future is only getting better, where the AI economy is headed, human relationships with AIs, longevity escape velocity, and so much more.

AI Futurist Ray Kurzweil invented the law of accelerating returns which states that the rate of progress is increasing at an exponential rate over time. If the past few months are any indication, this law is alive and well. In just the last few weeks, we've seen context windows expand by leaps and bound with [Gemini](#) and Supermaven. We've seen minute long videos thanks to developments like [Sora](#). We've seen LLM models reach new heights with [Claude 3](#). And developers are adopting AI at an ever increasing pace. From healthcare and bio to [government](#) to cybersecurity to consumer applications, no sector is immune to the power of AI.

Virtually every net new startup we see at Bessemer is using or leveraging AI in some capacity.



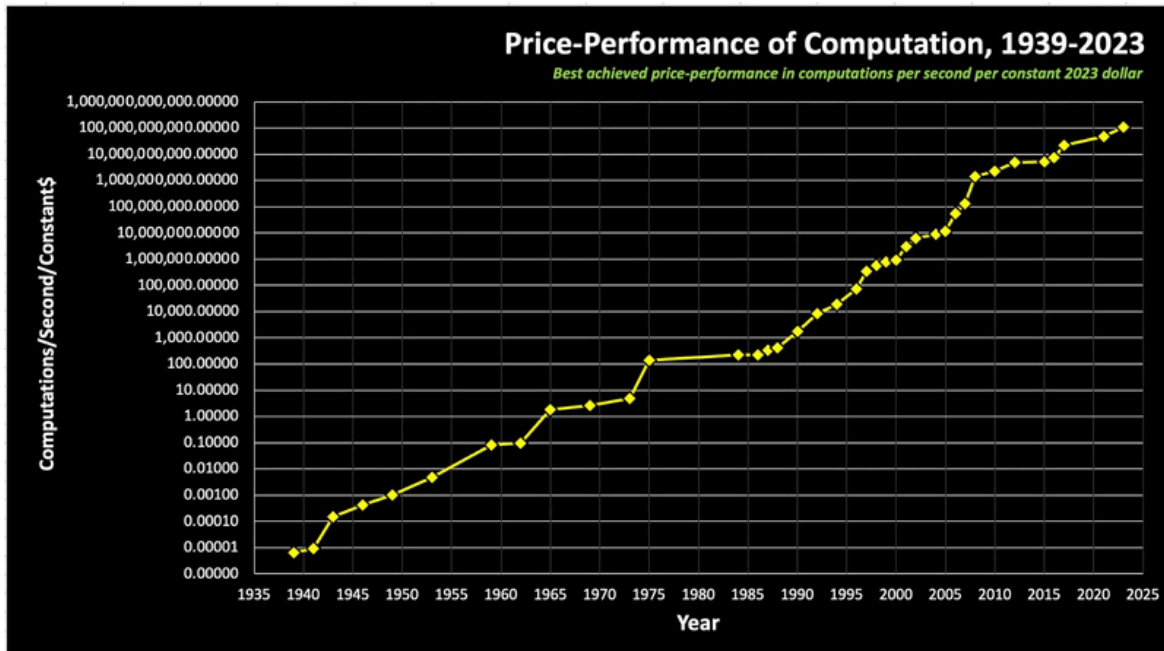
In our conversation with the one-and-only Dr. Ray Kurzweil, we explore the state of AI, and how life, technology, and well, everything could change in the next 10 years.

The power of timing

Talia Goldberg: I want to start by talking about timing. In the venture capital and startup industry, being too early is often the same as being wrong. And so, timing is everything.

Ray Kurzweil: Is that important in venture capital?

Talia: It really is everything, and it's everything for you too, as a futurist. It's about understanding how the world's going to evolve and what ideas can actually emerge at certain times. You've had a few big dates: 2029 for when we pass the Turing test, 2045 for the singularity. How do you think about timing? What leads you to these dates?



Ray: This is my main graph. This is not just about the present. This goes back to 1939. This is the power of computers per dollar, per constant dollar. It's a logarithmic graph. As you go up the graph, it's expanding exponentially. So this graph represents a 20 quadrillion fold increase for the amount of computation you can get for the same amount of money.

It started in 1931 with a German computer, which was actually presented to Hitler and he saw no reason to have computation and rejected it. It did a small fraction of a calculation. 0.000007 calculations per second per constant dollar.

We took the leading computer of that year—the latest one was actually a Google computer—and it's 130 billion calculations per dollar. So that's a 20 quadrillion fold increase.

But what's really amazing is that this went on for 40 years, nobody knew that it was happening. 40 years ago, I started looking at this and noticed that it was an absolute straight line. Regardless of what we were increasing—relay speeds or vacuum tube speeds or integrated circuits—it increased by the same amount every year for 80 years. Nobody even knew that it was happening for the first 40 years.

So, I thought for 40 years, for various reasons, it would continue. Maybe in wartime it would grow more quickly. But the rate is not affected by anything that's happening, including wars and so on. This is an example of the exponential rate of technology.

A lot [of the future] comes from this graph. We didn't have large language models 80 years ago, or even three years ago. But if you can figure out what is required in order to create [new innovations and technologies] you can predict [the timing of] various things.

In 1999 I projected the [AI advances] would continue at this pace. I figured we would pass the Turing test within 30 years by 2029. Stanford felt that was very alarming, and so they held an international conference, and AI experts came from all over the world. They felt I was very over optimistic.



They felt it would take a hundred years. I'm still saying 2029, and it turns out to be pessimistic. A lot of people are saying [we will pass the Turing test] by next year. Some think it's already happened. However, the Turing test is actually not very well defined. (Turing wrote an essay about it.) So, I figured people would say we're passing a Turing test, but it wouldn't be real until most agree that we're passing a Turing test. I think that'll start next year.

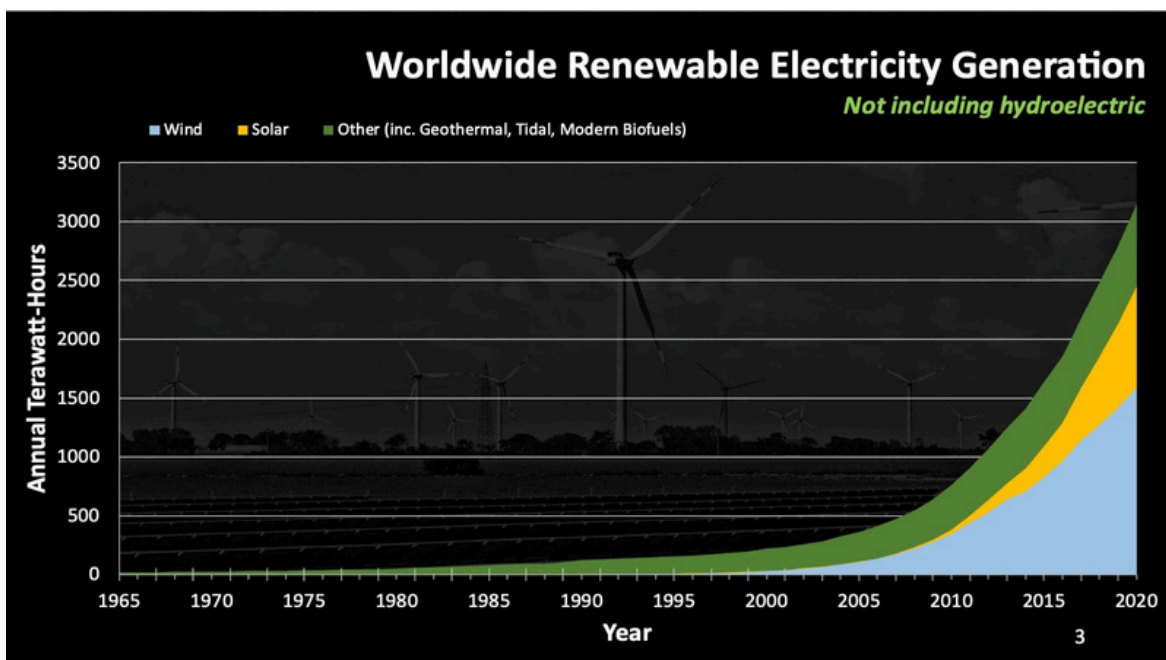
Our ability to create something that's super human is around the same speed. If you look at a large language model, there are certain things that an LLM can't do that humans can do. But there are lots of things an LLM can do that humans can't.

Take some obscure philosophy problem and say who can do this and it can write you a very intelligent essay about it and it takes about 20 seconds. No human being can do that. In fact, you can ask it anything and it can answer it pretty intelligently. No human being can do that.

[AI is] already going way beyond what a human being can do. This is what's driving technology. And this is just an example, a very important example. [The exponential growth of computation] is key to technology progressing.

Dynamics of exponential growth

Ray: We see [exponential advances] in every field. Here's renewable electricity generation:



This is an exponential rate, and it's going to pass a hundred percent within ten years.

Most economists have no idea that things progress in an exponential manner, and they're surprised that long term predictions are much more realistic because it takes into account the exponential. But most economists do not use an exponential at all. But this is exponential. It's going to pass 100%. And at that point, we'll be using one part in 10,000 of the sunlight that falls on the Earth. We have 10,000 times more sunlight to meet all of our energy needs than we need. We'll pass that within 10 years. I devised [these charts] to time my own projects. It was not originally intended to be a way to predict the future. But it provides a very good way to predict the future. In 1999, I predicted that we would have things like large language models now.

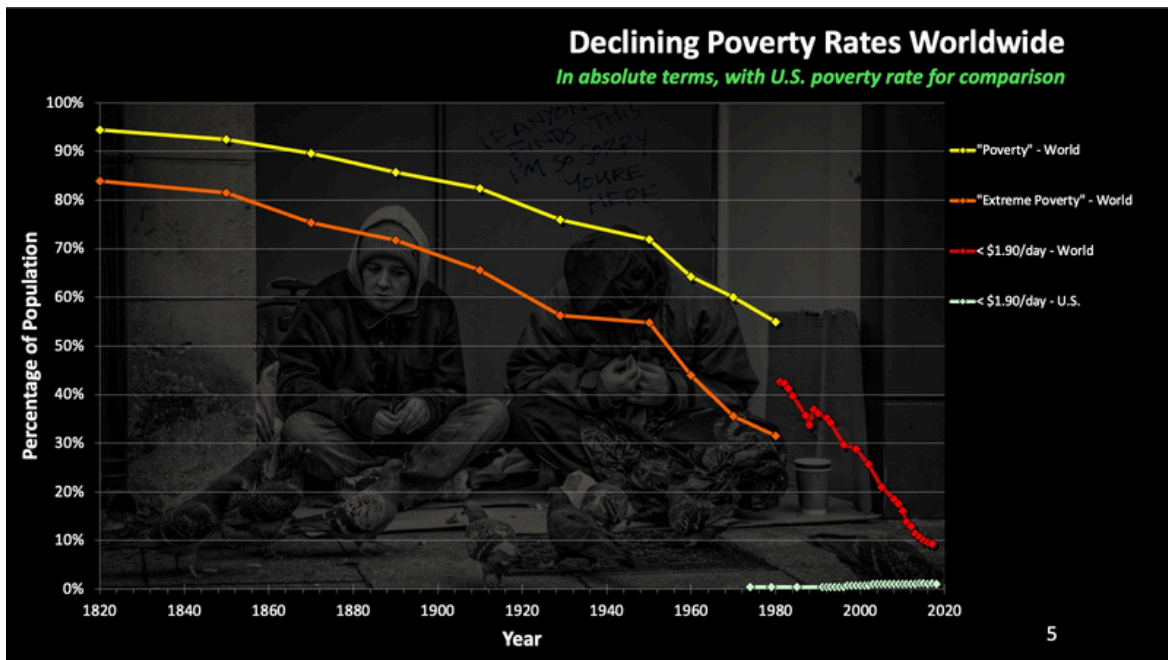


Why the future is getting better

Ray: People think things are actually getting worse. 31,000 people were asked: "Is poverty getting better or worse?" 66 percent felt it's actually getting worse. The reality is we've actually reduced poverty 50 percent over the last 20 years. And that was the answer of only 2 percent of the people that answered this.

"People tend to have a negative view of the future. They think things are getting worse. They are actually getting dramatically better."

I've got 50 different graphs [like this].

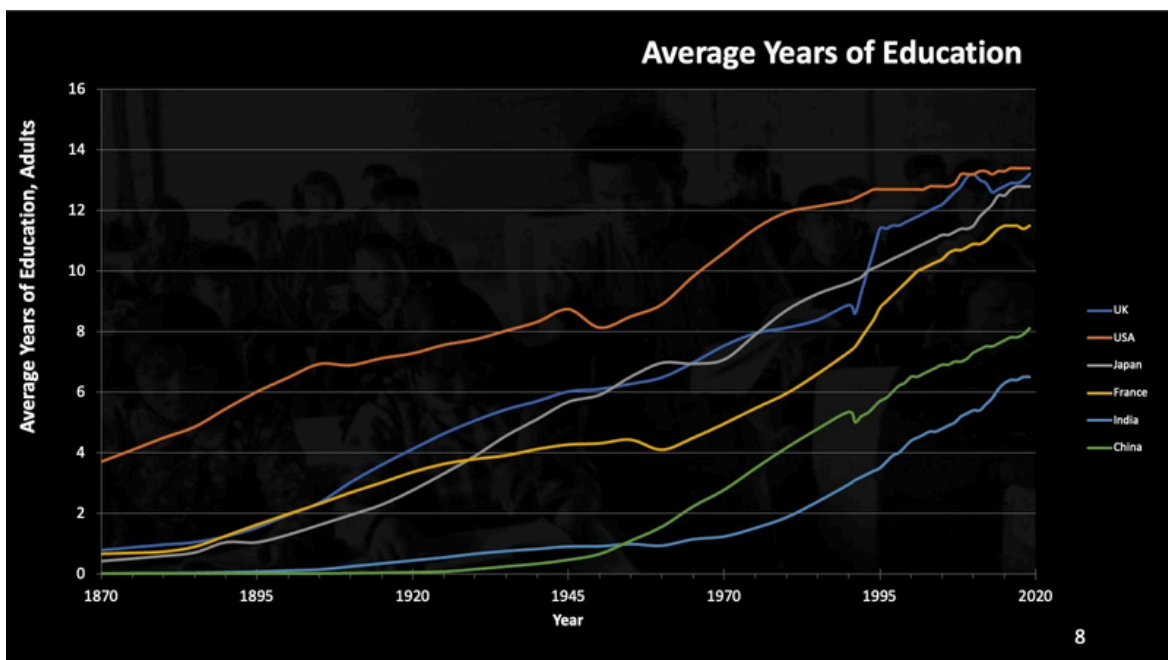
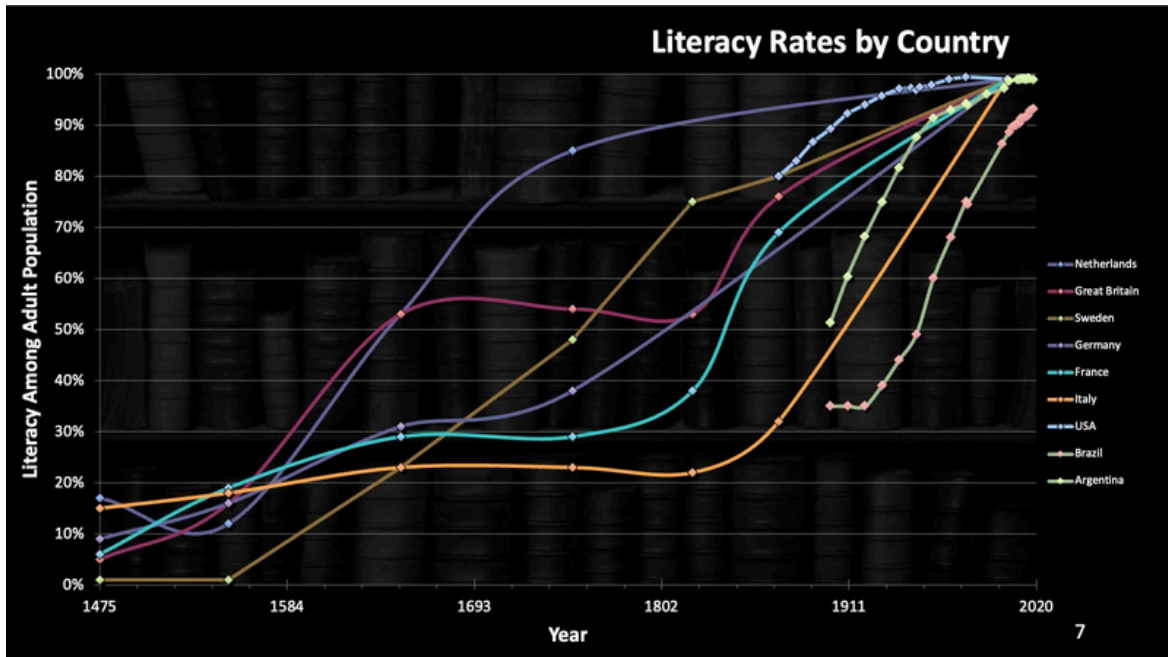


Declining poverty rates.





This is U.S. personal income. This is in constant dollars. This is the amount of money everybody has on average. It's fantastically soaring. If you go back a hundred years things were terrible. There were no anti-poverty programs that could account for that. And we, of course, didn't have a lot of the inventions that we have today. But even despite that, there was very little money to go around.



Literacy rates, averages of education, we have 50 different graphs.

Some have noticed [that things are indeed getting better] and have articulated this, but they have not connected [things getting better because of] the exponential growth of technology [and computing power.]



For example, I was trying to do optical character recognition ("OCR"), and I decided to do that in the 70s, because it doesn't require that much calculation because the lines [of text characters] are straight. I decided to do speech recognition in the 1980s because people's voices are much more different and it requires more calculations. So we invented something called Markov models, a little bit like neural nets today.

Those required a lot more computation. So I've used [the exponential growth of compute] to time my own projects. But now it provides a convenient way to predict when things will happen. That's what my books are about. I have a book coming out in May: The Singularity is Nearer. This is after my most famous book, The Singularity is Near, which is now 20 years old, but people still read it.

But economists still do not understand that technology progresses exponentially, which is amazing, because it's right there, and it really is animating everything we do. You're the ones that take advantage of that, because you can invest in things.

New opportunities that come along as technology speeds up. Mostly computers, but it's everything else as well.

AI breakthroughs

Talia Goldberg: Sam Altman has also thought about these exponential improvements [and appreciated the power of] scaling laws. GPT-2 wasn't very useful. But as [the models] scaled and they threw more compute and more data at it, the performance became remarkable.

Do you think that we have the right algorithms today to build general purpose human level AI? Is it just a matter of scaling up deep neural nets? Or is there like a fundamental breakthrough that needs to happen?

Ray Kurzweil: I think we have the calculations. We don't have the data. And the data is actually very important and will become the key issue. The models we have now actually have enough capacity, but we don't have the data. And so finding ways to get data that can handle these very large models is going to be key.

You need a trillion parameters. GPT 4 is probably 400 billion. So it's not quite there. I can't exactly tell you what the Google ones are, but we're not really at a trillion. But I think a few trillion parameters will be sufficient. But how you [feed these models] data that's accurate, [that is important]. Google just had this [challenge now with image generation] where the data [and outputs] are incorrect.

They're going to fix that. They better fix it. But having correct data is very important. Particularly if you're dealing with private data, not just public data that's where the value is going to be, to be able to actually characterize it with, with correct data.

Talia: It sounds like you think that to get to general intelligence we can just continue on the scaling laws. To get to the singularity, is that the case too?

Ray: There are various tricks that we find which change into algorithms. We're learning how to deal with this data, but we don't have sufficient algorithms that can handle the volume of data that we need. So we're developing algorithms and ways to create data that are much more efficient than what we have today. That's going to continue. It's happening every week.



What Ray thinks of Nvidia

Talia: Ray, you would be an amazing investor. I hope that you've been investing in the stock market, because as I was preparing for this, I saw that in 2005, in the Singularity is Near, you were one of the first people to realize that gaming GPUs should run on neural networks. NVIDIA could become the largest company in the world over the next few years. What do you think is the future for NVIDIA?

Ray: I mean, it looks good! But whether people are overly invested, it's hard to say. They're not the only people that have that algorithm. In fact, the last two charts I showed are actually Google chips, which outperform NVIDIA chips. [Nvidia is] very well positioned, but we didn't hear about them ten years ago, and other people may overtake them. They have a good position, but the stock market is evaluating them to have a good position. They're not the only ones to have this kind of chip that can run large language models.

Open vs. Closed Players

Talia Goldberg: One thing that's so amazing about this chart, and one of the reasons I really like it, is that it's not just about performance, but also cost.

At Bessemer, we're huge believers and advocates of open source AI. We're funding a lot of initiatives in and around the open source AI ecosystem. What's exciting is that what's possible in some of the largest labs at Google increasingly is now possible in someone's garage. It's becoming democratized. It's getting cheaper. It's more accessible. It's part of what's leading to the chart that Bhavik showed with more and more startups emerging. It's really important for our business. [Open source and accessible AI] is really important for humanity. There's a tension between closed AI and truly open AI and the open source AI ecosystem.

How do you think the rate of progress in the future changes if we have a world that's dominated by a few, concentrated closed models and players — Google and OpenAI and the like — versus a much more fragmented, open ecosystem?

Ray Kurzweil: There's no reason for that to happen. We have both today. And they really serve different purposes. And they're both going to have market applications.

The closed model might be very good for a private data set that only one person controls. And they can feed it into their own private data bank. So I think both are going to coexist.

The public ones are the most impactful because everybody can use them and it uses things that we all acknowledge as being the same, but it can be vast and we're constantly increasing the size of them. The public models will be the most potent, but closed models are also going to continue.

Talia: You think the public, meaning the open source ones will be, even though the closed ones can afford to put more and more money behind it?

Ray: Yeah.

Human-AI relationships

Talia Goldberg: Before this chat, I was talking with Ray about human-AI relationships. This is an area that has proven to have strong product market fit for large language models over the past couple of years. A lot of people don't like to talk about it, but it's very real. There are a billion lonely people or more in the world, and we already see companies like [Character AI](#) catering to that.



Ray Kurzweil: We see that in movies, R2-D2 and so on. Totally. They're machines, but we do treat them like pets.

Talia: What does that mean in terms of a future of dating AIs and having relationships with AIs? Should there be a distinction between a human relationship and a human to AI relationship?

Ray: It really depends on whether the AI has all the capabilities of a human. They're not quite there yet. I mean, this is not 2029. It's only 2024. We're gonna get there pretty quickly once AI has all the capabilities of a human and can speak like a human — not just sequence words like a human. AIs have to care about people and can actually be even more insightful into human's motivations than humans are today. There's no reason not to have a relationship with AIs.

Talia: Except — well, except for the physical.

Ray: Eventually we'll be able to match that as well. AI-based humans are coming. They're not quite at the level of humans, but they will be. They'll be able to move just like a human. It's not today. I'm involved with one company [that is building an AI robot that has] fingers that move just like a human, it can open a bottle — it's coming.

But once AIs actually have all the physical capabilities of humans and a mind like a human, and can understand humans, it'll go beyond what humans can do today. It might be preferable to have a relationship with an AI.

Talia: So in this future world, would you be comfortable with your grandchildren having their partner be an AI? Ethan, don't answer that.

Ray: We'll have relationships with AIs... depends on what your ideas are about relationships. But we can have a very beneficial relationship with AI. That's definitely going to happen.

Talia: An AI could be an infinitely supportive and empathetic partner. Based on the consumer usage data we track for large language models and AI applications, intimate, inherently human relationship and dialogue based connections are taking off. It's coming fast whether we like it or not. Society has to acknowledge that.

Humans merging with AI

Talia Goldberg: There's another idea that you have that I want to talk about: humans merging with AI. I already see this happening. For example, my cell phone is already an extension of me, to some degree. I don't leave my house without it. It augments my life. It tells me where I am, what's going on.

We can transcend the limitations of biology. You've written a lot about amplifying our brain with technology and AI. When do you think that happens? How do we actually transcend the limits of the biological body and brain so humans can merge with AI?

We see things like [Neuralink](#). I'd be interested in your thoughts on Neuralink. And if you're timing the future and predicting the human brain merging with AI, where do you place that in time?

Ray Kurzweil: Let me comment on Neuralink. It isn't a predecessor of being able to interact with a neocortex. Neuralink itself is very slow. The application of it is really to provide a method of communication for people that can't communicate. I know some people who were scientists, who used to be very vocal but now cannot communicate at all. And so we're looking at using things like Neuralink to enable them to communicate.

There was one person who became "locked in" and they gave him one of these implants, and his first word was [antidisestablishmentarianism](#). He said that to show that he was thinking. How amazing. Because people thought that maybe he had lost his cognitive ability. But he hadn't. That's what things like Neuralink are for.



Talia: When will we be able to connect our neocortex to compute at a sufficiently high bandwidth?

Ray: Early 2030s. At that time you can have a human being that has the entire capacity of a large language model inside their brain.

Talia: Kinda like cloud computing, but having your brain connected to the large language model?

Ray: People will say, "I don't think I would want that." But it's like saying "I don't think I'd want to use a phone!" Who here doesn't have their phone? Probably nobody. Ten years ago nobody had their phones. Now everybody has their phones. They are an extension of our mind. But we're not connected directly to it so there's a certain amount of interaction between it and us that slows things down. If it actually was directly connected to our brains, we could have all that capacity and information inside our brains.

Technology is an extender of human thought. People are very concerned about us versus AI, as if it is an intelligence that comes from another planet. But it's created by human beings. It's based on human thought and it amplifies who we are. We should be enthusiastic about it because it's amplifying who human beings are.

Talia: We're lucky to be alive right now. The future is exciting.

Longevity escape velocity

Talia Goldberg: I want to talk about longevity escape velocity. It's a topic I'm passionate about, and you've been talking about it for a long time.

For those that aren't aware, longevity escape velocity is when life expectancy increases by more than a year per year.

Ray Kurzweil: Right. So right now you go through a year and use up a year of your longevity. However, research is advancing and it's curing various diseases. You're actually getting back on average about four months a year. So you lose a year of longevity. You get back about four months because of scientific research. However, scientific research is also on an exponential curve. By 2029, you'll get back a full year. So you lose a year, but you get back a year.

Past 2029, you'll get back more than a year. Go backwards in time. Once you can get back at least a year, you've reached longevity escape velocity.

Talia: So it's safe to say that you think everyone in this room can reach longevity escape velocity.

Ray: Yes. Now that doesn't guarantee you living forever. You could have a 10 year old and you could compute that he's got many, many decades of longevity, but he could die tomorrow.

Although we're also addressing things like accidents. Self driving cars will virtually eliminate many accidents. We're not quite there yet. But that will happen as well.

Talia: Does longevity escape velocity apply to the body too? Is the body going to decay? Or will we upgrade our body parts at various points? How do we physically get to longevity escape velocity?

Ray: We're constantly replacing a lot of our body all the time. In fact, most of our body is constantly dying and recreating. And as we go through more and more scientific progress, we'll extend that. Forever. So once you're past longevity escape velocity, you go through a year, you're not a year older and you're not more likely to die.

You can be less likely to die. Doesn't guarantee that you won't have an accident, but that's really where we're headed.



Ray: That's an old statistic. We're constantly speeding that up. Moderna, for example, created its vaccine in two days. It got out in a total of ten months. So that 19 years stat is increasingly outdated. There's a slow part of medicine, but it is already greatly sped up.

We got the COVID vaccine out in ten months. It took two days to create it. Because we sequenced through several billion different mRNA sequences in two days. There's many other advances happening. We're starting to see simulated biology being used and that's one of the reasons that we're going to make so much progress in the next five years.

Cryonics

Talia: If you were to face death before reaching longevity escape velocity, would you consider cryonics? (For those unaware, cryonics is where you use liquid nitrogen to freeze and preserve the corpse and the brain with the hopes that then we can reanimate and regenerate in the future.)

Ray: Yes. I'm signed up for it. If you're going to die, there's no alternative. For those that are burnt after death or so on, they obviously have no ways of reanimating their bodies. Cryonics just gives you a possibility. It's quite feasible to see how that could happen in 20, 30, 40 years. But it's not guaranteed. My real strategy is to reach longevity escape velocity, and not die.

Talia: Any tips?

Ray: I've written three books on how to stay healthy. I take about 80 pills a day, different injections, and so on. I'll just mention one: Lipitor. It both reduces your LDL. My LDL is like 25. And my HDL was like 25, but now it's like 50. So that's one pill that's pretty effective.

Advice for the future

Talia: The rate of progress is changing exponentially, or improving exponentially. What would you impart on the next generation of humans and the entrepreneurs that are here building and creating and inventing the future?

Ray: I've tried to foster an idea of innovation for both my children.

My daughter [Amy Kurzweil](#) is a noted cartoonist and almost has a record for the number of cartoons in the New Yorker. So she's very creative and constantly looks at different ways that society can manifest some kind of whimsy. And [Ethan](#)'s done the same thing. Venture capital is not one method that solves all problems — it's constantly changing. Not only what you invest in, but how you invest, and how you create relationships, and how you create social networks. I've been tracking venture capital going back now for 50, 60 years and it's the way in which it's done is constantly on the cutting edge.

Will AI have free will?

Ray Kurzweil: In my new book, "Singularity is Nearer," I explore the concept of free will. The future is deterministic but ultimately unknowable until it unfolds. We can't predict it or simulate it in advance. Similarly, AI will possess free will in the sense that their actions are unpredictable until observed.

Talia Goldberg: It sounds like you consider humans deterministic as well.

Ray: Yes, our actions are determined, but they're unpredictable. It's akin to Wolfram's Stage Four complexity—you can't determine the outcome without actually going through the process.



What about the limits to growth? The nice straight line on the graph you showed will eventually flatten out, right?

Ray: Nanotechnology could create computers far beyond today's capabilities. A one-liter device could perform as many calculations as all humans combined. The limits aren't foreseeable yet; the growth will continue for a long time.

Audience: But it will level off eventually?

Ray: Perhaps, but it will be so far beyond our current comprehension that it's hard to speculate on the cause.

Audience: What about population growth? When will humanity reach its capacity?

Ray: We might not reach the projected 10 billion due to declining birth rates. Intelligence will increasingly come from AI, which will be rooted in but extend beyond human intelligence. The unique aspect of humans is our ability to create technology, which comes from our dexterity, exemplified by the thumb. Other intelligent species can't do this.

In a post-singularity world, what is the relevance of individual humans?

Ray: Human beings and technology will merge. The fusion is what's important—not the separation. Technology is created to solve our problems and augment our capabilities.

Will AI become a separate social entity with rights?

Ray: It's hard to say if that will happen. Technology is an extension of humans and stems from us. We create technology, which is unique to our species.

Will you create a digital clone of your brain?

Ray: I've already done something similar with my father's writings. I'd like to do the same with my own work. As for brain-computer interfaces, AGI will likely develop more quickly, but merging with AI is a natural progression we're already experiencing with current technology.

How does a futurist predict effectively?

It starts with understanding the exponential growth of computation. From there, it takes some imagination to envision future applications. This method is relatively new in futurism but fundamental to predicting technological advancements.

To listen to the full conversation between Ray Kurzweil and Talia Goldberg go to <https://www.bvp.com/atlas/ai-escape-velocity-a-conversation-with-ray-kurzweil>.

OUR \$1 BILLION COMMITMENT TO AI

Crossing The Rubicon



Venture insights that matter

bvp.com/subscribe



With Artificial Intelligence, there has never been a better time for small, ambitious founding teams to positively transform life as we know it.

We are amidst a major computing revolution. Artificial intelligence is here and nearing escape velocity. Progress across numerous technological vectors has led us to this point — from new model architectures to specialized hardware with vast computing power to advanced machine learning techniques.

There has never been a better time for small, ambitious teams to positively transform life as we know it. AI is no longer locked in research labs. The open source ecosystem is thriving. APIs are accessible. Costs are plummeting.

For consumers, AI is already permeating daily life. [ChatGPT](#) captured the imagination of over 100 million users practically overnight. [Midjourney](#) brings human imagination to life and supercharges creativity. Human-AI communication is becoming mainstream. We expect AI agents will soon outnumber humans.

For scientists, AI serves as a multiplier for transformational research and discovery, from life-saving drug discovery to personalized medicine ([Peptone](#), [Alphafold](#)) to AIs that comb satellite photos of the Earth for methane leaks and CT scans for tumors.

For businesses, AI improves efficiency while unlocking new customer experiences and opportunities. Just look at [Abridge](#), [Intercom](#), [Ada](#), or [Jasper](#). Products like [Copilot](#) imbue team members with superpowers. Platforms like [HuggingFace](#) and [Zapier](#) lower the barriers to innovation and broaden access to AI.

We have crossed the Rubicon.

With this in mind, Bessemer is committing \$1 billion in capital from its current funds to support entrepreneurs at the bleeding edges of AI and founders building AI-native products.

We aren't interested in faster horses, we want to invent automobiles. We aren't funding features or marginal improvement, we are looking for breakthroughs. And the breakouts will be built ground up with AI, by those that are agile, technical, and unencumbered by the status quo.

Our goal is to support the entrepreneurs forging new paths with AI to accelerate positive progress, transform markets, and imagine new ones.

Capital alone is not enough. We are building a community for AI founders to tap into a network of operating advisors, share field notes, and get preferred access to compute and models. We are also kicking off a series of workshops specific to companies building and innovating in AI.

And in the spirit of experimentation and learning, we launched [ChatBVP!](#) ChatBVP is your friendly neighborhood AI chatbot, here to support founders and answer questions about Bessemer. Please introduce yourself, have fun, and let us know what you think.

If you're building an AI-native startup, we want to support you with advice, community and capital. Reach out to us at ai@bvp.com and on bvp.com/ai.



Venture insights that matter

bvp.com/subscribe