The Data Bottleneck: Why Andreesse Horowitz Bet $30M on Protege

JAN 11, 2026

Table of Contents

Abstract

The AI industry faces a fundamental constraint that compute and architecture improvements cannot solve: access to high-quality, real-world data. Public datas have been exhausted, synthetic data has hit scaling limits, and the vast majority valuable data remains locked in private systems across healthcare, enterprises, a operational environments. Andreessen Horowitz's decision to lead a $30M Serie extension in Protege represents a thesis that data infrastructure will be as foundational to AI as cloud infrastructure was to SaaS. The company, founded ir by Travis May and Bobby Samuels, is building the platform that connects AI bui with massive, multimodal datasets across healthcare, video, audio, motion captu

and other domains. May previously founded and led both LiveRamp and Datavar major exits, while Samuels brings operational expertise from leadership roles at companies. This essay examines why a16z believes data access is the critical bottleneck for AI advancement, why this specific team has unique advantages to it, and what the emergence of functional data infrastructure means for AI development economics.

The Exhaustion Problem

The progression of language models from GPT-2 to GPT-4 and beyond tells a cle story about the role of data in AI advancement. Early gains came from better architectures and more compute. Transformers beat LSTMs. Scaling laws held. M parameters plus more GPUs equaled better performance. But somewhere around 2023, the easy wins from architecture and compute started running into a hard w Not because the models could not scale further, but because the training data co not.

Common Crawl has been scraped to death. Reddit threads from 2012 have been ingested a dozen times over. GitHub repositories are exhausted. Wikipedia exist every major model's training corpus. The entire public internet, which seemed in when these projects started, turns out to be finite and largely consumed. Synthet data generation helps at the margins but cannot replace real-world complexity. Models trained primarily on synthetic data tend to collapse into repetitive patter hallucinate in predictable ways when faced with novel situations.

The problem extends beyond text. Computer vision models need diverse, high-q labeled images and videos that capture edge cases, rare events, and unusual light conditions. Audio models need clean recordings across accents, environments, a acoustic conditions. Robotics and embodied AI need sensor data from physical environments. Medical AI needs patient outcomes across diverse populations an treatment contexts. All of this data exists, but almost none of it is publicly availa easily accessible.

Meanwhile, model architectures are converging. The difference between leading frontier models has less to do with fundamental architectural innovations and m do with training data quality, instruction tuning datasets, and RLHF approaches When Anthropic releases a new Claude variant or Google ships an updated Gem model, the competitive advantage often comes down to what data they trained or how they curated it, not whether they invented a novel attention mechanism.

This creates an uncomfortable reality for AI builders. The next 10x improvemen model capability will not come primarily from buying more H100s or hiring mor researchers. It will come from getting access to better training data. Specifically, world data that captures the messy, multimodal, high-stakes environments wher systems will actually operate. The internet represents maybe 5% of the world's to data. The other 95% sits in hospitals, enterprises, research labs, media archives, operational systems. Unlocking that data is the problem Protege is solving.

Why Travis May Built This Again

Travis May has spent nearly two decades building data infrastructure companies Protege represents his third major swing at solving data fragmentation problems track record is about as good as it gets in enterprise data, with two successful companies already under his belt before starting Protege at age 37.

May co-founded LiveRamp in 2011 with Auren Hoffman, initially joining as VP Product before becoming CEO. The company, originally called Rapleaf, built ide resolution infrastructure for marketing and advertising, becoming the dominant platform for how brands connected customer data across different systems while maintaining privacy. LiveRamp was acquired by Acxiom for $310M in 2014, later out as an independent public company in 2018, and at its peak was processing da connections for basically every major brand and publisher.

After LiveRamp, May reconnected with Vivek Ramaswamy, a friend from Harvar where they had co-founded Campus Venture Network together in 2007. Ramasw had founded Roivant Sciences and wanted to rebuild LiveRamp's data connectivi capabilities but for healthcare. In 2017, May co-founded Datavant with Ramasw

with Roivant providing initial funding and May making a significant personal investment. May served as CEO of Datavant from 2017 to 2022, building it into t leading healthcare data infrastructure company. Datavant merged with Ciox Hea a $7B transaction in 2021, creating what became the largest neutral health data ecosystem in the US, connecting over 2,000 hospitals, 15,000 clinics, and hundre other healthcare organizations. May stepped down as CEO after the merger, bec President and remaining on the board.

The pattern across both LiveRamp and Datavant was identical. Build technical infrastructure to solve hard data problems that everyone faces but nobody wants build themselves. Create a neutral platform that benefits from network effects a more parties connect to it. Focus on privacy, compliance, and trust as core produ features, not afterthoughts. Charge reasonable prices so you become infrastructu rather than a vendor. Scale to the point where you are embedded in critical work and switching becomes prohibitively expensive.

What May saw with AI was that the same fundamental problem was emerging, ju a different context and at potentially much larger scale. Companies had spent a decade building massive model architectures and acquiring GPUs, but now they hitting a wall on training data. The valuable data was fragmented, locked up, regulated, and hard to access. Every AI company was trying to solve this problen independently, negotiating one-off deals with data providers, building custom pipelines, and burning resources on undifferentiated heavy lifting. It looked exac like the problem LiveRamp solved for marketing data and Datavant solved for pa data.

The difference this time was that AI represented a much larger market opportur Marketing data infrastructure was a multi-billion dollar market. Healthcare data infrastructure was bigger, with Datavant reaching a $7B valuation. But AI traini data infrastructure could be an order of magnitude larger because it cuts across vertical and every geography. Every frontier AI lab needs this. Every AI applicati company needs this. Every enterprise building internal AI capabilities needs this total addressable market is basically the entire AI industry, which is projected to worth trillions.

May also recognized that the window to build this was limited. Data infrastructu businesses have strong first-mover advantages because they benefit from networ effects on both sides. The platform that signs up the most data suppliers first be more attractive to AI builders. The platform with the most AI builder demand becomes more attractive to data suppliers. Once that flywheel starts spinning, it becomes hard to displace. LiveRamp won because it got to critical mass first in marketing. Datavant won because it got to critical mass first in health data. Prot has the opportunity to win in AI data infrastructure, but only if they move fast.

In 2024, May co-founded Protege with Bobby Samuels, who he had worked with both LiveRamp and Datavant. Samuels had been at LiveRamp in partnerships ro from 2014 to 2016, then joined Datavant in 2020, eventually becoming General Manager of Privacy Hub before leaving in 2023. The two brought in Engy Ziedar Chief Scientific Officer and Richard Ho as CTO to round out the founding team deep technical and scientific expertise.

The a16z Investment Thesis

Andreessen Horowitz leading a $30M Series A extension in Protege in January 2(signals strong conviction that data infrastructure will be foundational to AI advancement. The financing expanded the company's initial $25M Series A from August 2025, bringing total funding to $65M since founding in 2024. Returning investors include Footwork, CRV, Bloomberg Beta, Flex Capital, and Shaper Cap

The thesis breaks down into several components. First, data access is genuinely limiting factor for AI advancement right now. a16z's portfolio companies across and machine learning are all running into the same problem. They need diverse, quality training data and cannot get it efficiently. Startups are burning millions business development to cobble together datasets. Even well-funded companies struggle to access the data they need at the speed AI development requires. This creates demand for infrastructure that solves the problem systematically.

Second, the market is massive and growing. AI is eating every industry, and ever application needs training data specific to its domain. Healthcare AI needs patie

data. Autonomous vehicles need driving data. Robotics needs sensor data from physical environments. Media companies need content libraries. The total mark training data could be larger than cloud computing because it cuts across every case.

Third, network effects create defensibility. Once Protege has relationships with hundreds of data suppliers and dozens of major AI companies, new entrants face enormous barriers. Data suppliers will not want to manage relationships with multiple platforms. AI builders will not want to integrate with multiple data sou when one platform gives them everything. The winner in this market could be w take-most, similar to how Snowflake dominated cloud data warehousing or how Databricks dominated data lakehouse architecture.

Fourth, the team has done this before. May built LiveRamp into a public compa Datavant into a $7B business. Samuels brings operational expertise from leaders roles at both companies, including managing Datavant's Privacy Hub which was central to the company's compliance infrastructure. They know how to build neu infrastructure platforms that scale across an ecosystem. They understand compl privacy, and trust at a deep level. They have relationships with data suppliers an know how to structure partnerships. This is not a team figuring it out for the fir time. They are executing a proven playbook in a new market.

Fifth, timing is critical and favorable right now. AI companies are desperate for training data as public datasets run out. Frontier labs are willing to pay substant amounts for unique datasets. Data suppliers are waking up to the value of their assets and looking for ways to monetize them. Regulatory frameworks around A training data are still forming, creating an opportunity to help shape norms and standards. The window to build dominant data infrastructure is open but will n open forever.

The investment came from a16z's Bio and Health team, with partners Daisy Wol Eva Steinman involved. This makes sense given Protege's initial focus on healthc data, though the platform has expanded into video, audio, and motion capture. T

Bio and Health team's involvement suggests a16z sees healthcare as the beachhe market but understands the platform will expand across verticals.

The $30M round size on top of a previous $25M suggests a16z expects Protege t quickly. This is not a seed investment in an unproven team testing product-mark It is a bet that the team can rapidly build supply and demand network effects be competitors emerge. The capital likely goes toward hiring engineers to build tec infrastructure, business development to sign data suppliers, sales to land AI con customers, and compliance infrastructure to operate across jurisdictions.

What Protege Actually Does

Protege operates as a two-sided marketplace connecting data suppliers with AI builders, but calling it a marketplace undersells the technical and operational complexity involved. On the supply side, Protege partners with hospitals, health systems, labs, imaging centers, research networks, media companies, and other d holders. According to the company's announcements, Protege expanded its data partner network to hundreds of organizations in 2025, providing aggregated acc new data sources and formats.

Each partnership involves negotiating data licensing terms, building technical integrations to extract and normalize data, implementing privacy and complianc controls, and establishing revenue sharing arrangements. Protege provides rever share payouts to data partners with each use, creating an economic incentive for holders to contribute to the platform.

For healthcare specifically, Protege securely obtains patient data from multiple sources and stitches it into longitudinal, multimodal, anonymized patient-level datasets. This requires sophisticated entity resolution to match patient records a facilities without using identifiable information. A patient might have records a different hospitals, two labs, and an imaging center, all under slightly different n spellings or with different identifiers. Protege's algorithms match these records probabilistically while maintaining HIPAA compliance through tokenization an other privacy-preserving techniques.

The data itself comes in wildly different formats. EHR data arrives as HL7 messa FHIR resources, or proprietary formats depending on the source system. Lab res use LOINC codes. Diagnoses use ICD-10. Medications use RxNorm. Imaging da lives in DICOM files. Clinical notes are unstructured text. Protege normalizes al this into consistent schemas and data models that AI companies can actually use training without building custom parsers for every data source.

Quality control happens at multiple stages. Protege validates data completeness, checks for anomalies, scores data quality, and flags potential issues before delive datasets to customers. Bad training data causes model failures that might not su until production, so quality assurance cannot be an afterthought. The platform t data lineage, versions datasets, and maintains audit trails for compliance purpos

On the demand side, Protege serves frontier AI labs, AI application companies, a enterprises building internal AI capabilities. According to a16z's announcement Protege already works with the majority of MAG7 public companies plus many l private AI players. These companies use Protege to access curated datasets acros healthcare, video, audio, motion capture, and other modalities without needing t negotiate hundreds of individual data partnerships.

The platform delivers data through multiple mechanisms depending on custome needs. Protege curates datasets from across its partner network to meet AI development needs, providing AI-ready data that integrates with modern ML workflows. The key value proposition is enabling AI builders to iterate quickly o model development rather than spending months or years on data acquisition an cleaning.

Beyond healthcare, Protege has expanded into other data modalities where simil problems exist. Media companies have vast archives of video and audio content t valuable for training multimodal AI models but difficult to license at scale. Mot capture data from sports, entertainment, and research applications can train rob and embodied AI systems. The same platform architecture that aggregates healt data can aggregate content libraries, with appropriate adjustments for different licensing and compliance requirements.

Why This Team Can Execute

The reason a16z bet on this specific team rather than waiting for alternatives is t
May, Samuels, and co-founders have unfair advantages that are nearly impossibl
replicate. The relationship network alone is worth years of business developmen
May spent six years at Datavant building partnerships with major health systems
networks, imaging center chains, and health data vendors. Samuels spent over th
years at Datavant in roles including General Manager of Privacy Hub, giving hin
relationships and operational knowledge. Before that, both had experience at
LiveRamp building data partnerships in a different vertical.

Protege gets to inherit much of that relationship capital. When May calls the CE
a major health system or lab network to discuss contributing data to AI training
not a stranger pitching a new idea. He is someone they worked with successfully
Datavant. The trust is already established. The credibility is already there. The
understanding of their business constraints and compliance requirements is alre
built. This accelerates partnership development by an order of magnitude compa
to a new team starting from scratch.

The technical expertise is equally important. Building infrastructure that can in
data from thousands of heterogeneous sources, normalize it, maintain quality,
preserve privacy, and deliver it at scale is genuinely hard. Most startups would s
years just building the technical foundation before delivering value to customers
Datavant playbook, which May developed and Samuels helped execute, provides
blueprint for much of this. The core algorithms for entity resolution, the privacy
preserving techniques for tokenization, the data normalization pipelines, and th
compliance infrastructure all transfer to Protege with modifications.

The regulatory and compliance knowledge is maybe the most underrated advant
Healthcare data is among the most heavily regulated in the world. HIPAA has co
requirements around de-identification, business associate agreements, breach
notification, and audit trails. Different states have additional privacy laws.
International markets have GDPR and other frameworks. May and Samuels have
years working with healthcare lawyers, privacy officers, compliance teams, and

regulators. They know what is permissible, what requires special handling, and h
structure agreements that satisfy all parties.

This expertise extends to understanding how different data holders think about
sharing. Academic medical centers have research missions and IRB processes.
Community hospitals care about revenue and liability. Labs want to protect
competitive advantages. Health systems worry about patient trust. Protege can
structure partnerships that address each institution's specific concerns because
team has done it hundreds of times before at Datavant.

The go-to-market approach also benefits from proven patterns. Datavant succee
becoming neutral infrastructure that everyone could use without creating compe
disadvantages. Health systems shared data through Datavant because it did not f
any single pharma company or technology platform. The same neutrality princip
applies at Protege. AI companies will share a data platform if it does not give
preferential treatment to competitors. Data suppliers will contribute to a platfo
that treats all customers fairly.

Operationally, the team knows how to scale quickly while maintaining quality.
Datavant grew from zero to $7B valuation in under four years through its merge
Ciox in 2021. The playbook involves starting with a wedge use case that delivers
immediate value, proving the platform works, then expanding horizontally into
adjacent datasets and vertically into new customer segments. Healthcare was the
wedge for Protege. Video, audio, and motion capture are the horizontal expansio
Foundation model labs vs application companies vs enterprises represent vertica
segmentation.

The engineering talent required to build this platform is also easier to recruit w
the founders have successful exits and track records. Top data engineers want to
on hard problems with teams that have proven they can execute. Protege can attr
senior technical talent from companies like Databricks, Snowflake, and Palantir
offering equity in a rocket ship with experienced founders who have built
infrastructure companies before.

Economic Realignment

The emergence of Protege and similar data infrastructure platforms shifts econc throughout the AI stack in ways that are still playing out. For data suppliers, it c new revenue streams that never existed before. Hospitals and health systems hav always viewed patient data as a compliance burden and liability, not an asset. EM systems cost millions to maintain, data teams prevent breaches, and sharing data opens up risk. But if you can monetize anonymized data for AI training while maintaining full compliance, suddenly that liability becomes valuable.

For healthcare providers specifically, the economics are compelling. A mid-size hospital system sitting on ten years of EHR data, imaging, and lab results repres significant value for training diagnostic models or clinical decision support syste Previously, accessing that value required building internal data science teams, negotiating one-off partnerships, or simply leaving money on the table. Platform Protege that handle acquisition, anonymization, and licensing let providers gene revenue without adding headcount or compliance risk.

The revenue potential is meaningful relative to hospital margins. Health systems operate on thin margins, often 2 to 3% for non-profit hospitals. Adding a new re stream from data licensing, even if modest, can impact financial performance meaningfully. For struggling rural hospitals or safety-net providers, this could b difference between staying open and closing.

Research networks and registries face similar dynamics. Organizations that colle patient outcomes data for specific conditions or treatments have spent years bui these datasets for academic research. Now they can make that data available for development with appropriate protections, creating funding that makes their co research mission more sustainable. Disease-specific registries, tumor boards, an clinical trial networks all sit on valuable longitudinal outcome data that AI comp desperately need.

Media companies and content owners are waking up to similar opportunities. M studios and broadcasters have massive video and audio archives that were previo

just sitting in vaults or used for limited internal purposes. Training multimodal models on diverse video content has enormous value for companies building con vision, video generation, or embodied AI systems. Licensing historical content fo training creates a new revenue stream from otherwise dormant assets.

For AI builders, the economics flip from a major cost and bottleneck to a predic expense. Instead of hiring business development teams to negotiate dozens of hc partnerships, burning six to twelve months on each, companies can access curate datasets through Protege in weeks. Instead of building internal data engineering teams to clean and integrate heterogeneous sources, they get normalized data re for training. The time and cost savings are substantial, but the strategic value is larger.

Being able to iterate quickly on model hypotheses changes product development fundamentally. If you think adding a specific type of imaging data will improve diagnostic accuracy, you can test that in weeks rather than months or years. If a performs poorly on certain patient populations, you can quickly source addition training data to address the gap. Speed of iteration becomes a competitive advan and Protege enables that speed.

Pricing models will be critical for how this plays out. Traditional enterprise data involve lengthy negotiations, volume commitments, and opaque pricing. That wo for established companies with data budgets but kills startup experimentation. I Protege can offer transparent, usage-based pricing aligned to startup economics enables a much broader set of AI builders to access valuable training data. This similar to how AWS democratized infrastructure access compared to buying you servers.

There are interesting dynamics around data exclusivity and competitive advanta Should leading AI companies be able to license exclusive access to certain datase Does that create unfair advantages, or is it just normal competitive tactics? Prot needs to balance enabling competition with allowing differentiation. The likely equilibrium involves a mix of widely available datasets that level the playing fiel

exclusive arrangements for unique data sources, similar to how cloud infrastruct works today.

The revenue split between Protege and data suppliers also matters. If Protege ta too much margin, data suppliers will try to go direct or use competing platforms Protege gives away too much margin, the business will not be sustainable or pro enough to justify a16z's valuation expectations. The right split probably varies b type, exclusivity, and supplier bargaining power. Large health systems have more leverage than small research networks. Unique datasets command better econon than commoditized data.

What This Means for Builders

For founders building AI companies, the implications of mature data infrastruct shift strategic priorities in several ways. Data strategy moves from being primari business development and operations challenge to being a product and engineer question. Instead of hiring salespeople to negotiate hospital partnerships, you hi ML engineers to evaluate dataset quality and design training pipelines. Instead o building custom ETL for each data source, you integrate with standardized APIs

This lowers barriers to entry for new AI applications that were previously too di for startups to pursue. Building a diagnostic radiology model used to require yea partnerships before training the first model. Now you can get started in weeks. T opens up entire categories of healthcare AI that were only accessible to well-fun experienced teams. The same pattern will play out in other verticals as Protege a similar platforms expand beyond healthcare.

Competitive dynamics shift toward model architecture, training techniques, and application-specific optimization rather than pure data access. When everyone c access similar baseline datasets, differentiation comes from what you do with th This is probably healthier for innovation overall, since it rewards technical capa rather than just partnership skills. Companies compete on actual AI capabilities instead of who negotiated better data deals.

For investors, data infrastructure platforms represent a different risk-return pro than typical SaaS businesses. Network effects are strong once you have critical m on both supply and demand sides. Marginal costs for incremental data sources a customers are relatively low compared to initial platform development. Switchin costs are moderate to high once AI companies integrate data pipelines into train workflows. The business model looks more like marketplace economics than traditional software.

Revenue concentration around a small number of large AI customers creates risl also validates product-market fit. If the leading frontier model builders all use y platform, that proves the value proposition strongly. The question becomes whe you can expand beyond anchor customers to serve the long tail of AI builders. Pr already working with majority of MAG7 companies plus large private AI players suggests they have the anchor customers locked in. Expanding to mid-market an smaller companies will determine ultimate market size.

Regulatory risk is material but manageable with proper compliance infrastructu Healthcare data in particular faces ongoing regulatory scrutiny, and rules aroun training data are still evolving. Platforms with deep compliance expertise and relationships with regulators have advantages, but everyone operating in this spa accepts some regulatory uncertainty. The team's Datavant experience navigating HIPAA, state privacy laws, and international frameworks is a significant deriskin factor for investors.

International expansion creates obvious growth opportunities but requires rebu supply-side relationships and compliance infrastructure in each major market. D localization requirements, different privacy regimes, and varying healthcare syst structures mean you cannot just flip a switch to operate in Europe or Asia. Comp that can execute international expansion effectively will build significant moats. and Samuels did this at Datavant, giving them playbooks for how to approach different markets.

The broader pattern extends beyond healthcare to any domain with valuable priv data. Manufacturing and industrial companies with sensor data from physical

processes could enable embodied AI and robotics. Financial institutions with transaction data could train better fraud detection and risk models. Telecommunications companies with network data could improve infrastructure optimization. The playbook established in healthcare likely applies across multi verticals, each of which could be as large as healthcare alone.

What remains uncertain is whether data infrastructure becomes a winner-take-r market or supports multiple specialized platforms. Arguments exist on both sid Network effects and economies of scale in building supply relationships favor concentration. But vertical specialization, regional focus, and different data modalities might support multiple winners. Healthcare alone might sustain seve platforms focusing on different data types or customer segments. The next few y will determine market structure.

The timing question matters significantly. Data infrastructure platforms that establish themselves now, while frontier AI labs are desperate for training data, be sticky even as the market matures. Companies that wait risk entering a mark with established incumbents and locked-up supply relationships. For entrepren the window to build in this category is open but probably measured in quarters, years. Protege's $65M in funding and a16z backing will accelerate their timeline make it harder for followers to catch up.

Ultimately, the shift from compute and architecture to data as the primary const in AI advancement represents a fundamental market transition. The companies build infrastructure to make real-world data accessible at scale are not just servi providers but enablers of the entire next wave of AI applications. That is rare positioning and, if executed well, creates durable value. Travis May has built thi of infrastructure company twice before with LiveRamp and Datavant. The a16z l that he and Bobby Samuels can do it again in a market that is potentially larger t their previous two companies combined.

4 Likes

← Previous

Nex

Discussion about this post

Comments Restacks

Write a comment...