VIVAMEDA RESEARCH

The Unicorn Signature

How future unicorns concentrate in top score bands before scaling

ABSTRACT

Vivameda predicts which companies are structurally ready to enter hypergrowth. The top 1% of companies our model ranks experience hypergrowth at 27x the base rate.

We tested the model against 22 historical tech unicorns at their pre-scaling year. 82% landed in the top 30% of scores. Of 6 unicorns held entirely out of training, 100% landed in the top 30%.

Built on the Vivameda Longitudinal Workforce Panel: 4 million companies, 48 million observations, 1950 to 2020.

TRY IT NOW - SAMPLE RAG FILE

Ingest the companion sample into ChatGPT or Claude in 5 minutes.

vivameda.com/Vivameda_Unicorn_Signature_RAG_Sample.md

Methodology

Vivameda Longitudinal Workforce Panel, 4M+ companies, 1950 to 2020
Time-safe features only. Company-level train/test split.
Gradient boosted tree model. AUC 0.889 on held-out companies.

Authors

Vivameda Research · oli@vivameda.com · vivameda.com

Introduction

Most company scoring systems see growth after it is already visible. By the time a headcount curve bends upward sharply, the company is on the radar of every investor, platform, and competitor in its market. The question for anyone screening for the next wave of scaling companies is not whether growth is happening. It is which companies are structurally ready to scale before the curve becomes obvious.

This paper presents a model that ranks companies by their likelihood of entering hypergrowth within the next 3 years, using only data observable at the time. We train on 28,664 companies and test on 12,285 the model never saw. We then evaluate the model against a 50-company cohort of named tech unicorns to ask a simple question: did these companies look different from peers at the year before they scaled? They did, and the model captures the difference.

We target hypergrowth rather than unicorn status because valuation events are externally reported, delayed, and unevenly captured. Workforce hypergrowth is observable directly and serves as a measurable scaling regime. The unicorn cohort is the validation showcase, not the training target.

WHY THIS MATTERS

Growth ranks. Workforce structure explains.

Headcount and growth tell you what is moving. Workforce structure and capability composition help an AI agent reason about whether an organization is structurally ready for the move. Vivameda packages both into a pre-computed signal file that AI agents can ingest directly.

FOUR FINDINGS

1. The model predicts hypergrowth on held-out companies.
Top 1% of scored companies hypergrow at 27x the base rate. Top 1% of test observations captured 27% of all positive hypergrowth outcomes in the held-out set. AUC 0.889.

2. Workforce structure carries independent signal.
A model using only workforce composition, with no growth, size, industry, or age data, still predicts hypergrowth at 4.4x the base rate at top 1%. The signal exists independently and serves as an interpretable explanation layer.

3. Historical unicorns concentrate in top scores.
Of 22 unicorns with a visible pre-scaling year, 82% landed in the top 30% of scores. Of 6 unicorns held entirely out of training, 100% landed in the top 30%.

4. The signal is observable years before scaling.
Our prediction window is 3 years. The unicorn validation scores companies at the year before their early-scaling pattern fires. The signature exists in workforce structure before headcount expansion becomes visible.

Methodology

Prediction target

We define the prediction target as hypergrowth: a year-over-year workforce growth rate above 100% within the next 3 years. We chose this target rather than unicorn status because valuation events are externally reported, delayed, and unevenly captured across regions and sectors. Workforce hypergrowth is observable directly inside the panel and functions as a clean measurable scaling regime. The test-set base rate is 0.41%.

Dataset

We use 41,000 companies from the Vivameda panel in tech and tech-adjacent industries (computer software, internet, IT services, computer & network security, financial services, real estate, computer games, design, automotive, logistics, health & wellness, entertainment, media production, broadcast media, leisure & travel). Each company appears for multiple years, producing 302,000 company-year observations across 2001 to 2016.

Train/test split and leakage controls

We split the dataset by company, not by year. 70% of companies are used for training (28,664 companies, 211,736 observations). The remaining 30% are held out for testing (12,285 companies, 90,233 observations). No company appears in both sets. Company identifiers are not used as features. All features for any given observation are computed using only data available at or before year T. This structure prevents both time-leakage and company-level memorization.

Features

Features include observed headcount, year-over-year growth rate, previous-year growth, growth acceleration, role diversity (distinct role buckets), role coverage, capability composition (which of 10 capabilities appear in the top 3), capability coverage, primary role concentration, industry, and company age.

Model and evaluation

We compare four baselines against the full model. Each is a gradient boosted tree ensemble (200 trees, max depth 3, learning rate 0.05) trained on the training set and evaluated on the held-out test set. The test set contains 366 positive hypergrowth outcomes out of 90,233 observations.

FINDING 1

The model predicts hypergrowth

On 12,285 companies the model never saw during training, the full model concentrates hypergrowth at 27x the base rate at top 1%.

The training set consists of 28,664 companies and 211,736 company-year observations from 2001 to 2016. The held-out test set consists of 12,285 different companies and 90,233 observations. No company appears in both sets. All features are time-safe.

Test-set base rate for hypergrowth within 3 years is 0.41%. The full model achieves AUC 0.889. At top 1% of scores, precision is 11.1% (27x lift). At top 5%, precision is 4.5% (11x lift). At top 10%, precision is 2.8% (7x lift).

A baseline using only headcount achieves AUC 0.536. Adding growth features reaches 0.884. Adding industry and company age reaches 0.891. Adding workforce-structure features produces the full model at AUC 0.889. The strongest baseline slightly outperforms the full model on AUC; workforce features still show standalone predictive signal and provide interpretability and enrichment value.

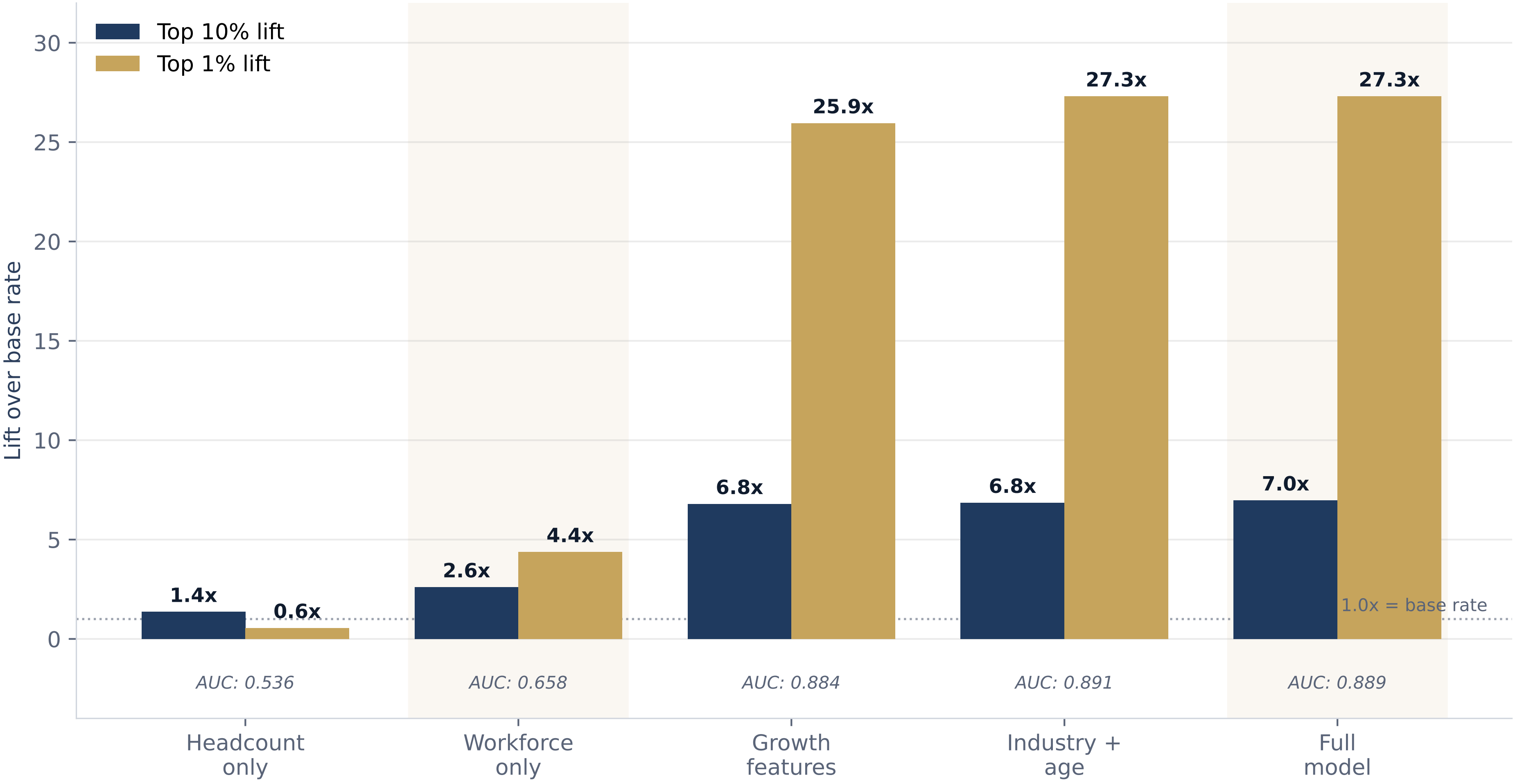
OPERATIONAL METRIC

1% of reviewed observations captures 27% of all hypergrowth.

Top 1% of the test set contains 902 observations. 100 of those experienced hypergrowth, out of 366 hypergrowth events total in the held-out set. Reviewing 1% of the universe captures 27% of the eventual scalers.

Model performance with ablation

Workforce features alone deliver real signal (4.4x at top 1%). The full model is dominated by growth and demographic features.



Source: Vivameda Workforce Panel · 43,208 companies, time-safe features, company-level train/test split

FINDING 2

Workforce structure carries independent signal

Growth ranks. Workforce explains. A model with no growth, size, industry, or age data still predicts hypergrowth at 4.4x the base rate at top 1%.

To isolate the contribution of workforce features, we trained a model using only capability composition, role diversity, role coverage, primary-role concentration, and capability coverage. No growth rate. No headcount. No industry. No company age.

This workforce-only model achieves AUC 0.658 on the held-out test set, with a top-1% lift of 4.4x and a top-10% lift of 2.6x. The signal is weaker than the full model but real and independent of demographic features.

In the full model, much of the workforce signal overlaps with growth and demographic features. The commercial value of the workforce layer is therefore not as a wholesale replacement for growth-based scoring. It is as an enrichment and explanation layer. Growth tells you what is moving; workforce structure helps an AI agent reason about whether the organization is structurally ready for that move.

FINDING 3

Historical unicorns concentrate in top scores

Of 22 unicorns with a visible pre-scaling year, 82% landed in the top 30% of scores. Of 6 unicorns held entirely out of training, 100% landed in the top 30%.

We scored each of the 50 unicorns in our validation cohort at their pre-scaling year, defined as the year before the early-scaling pattern first fired. 22 of the 50 had a pre-scaling year visible in the scoring window. The remaining 28 either entered the panel already in scaling state or did not have the pattern fire.

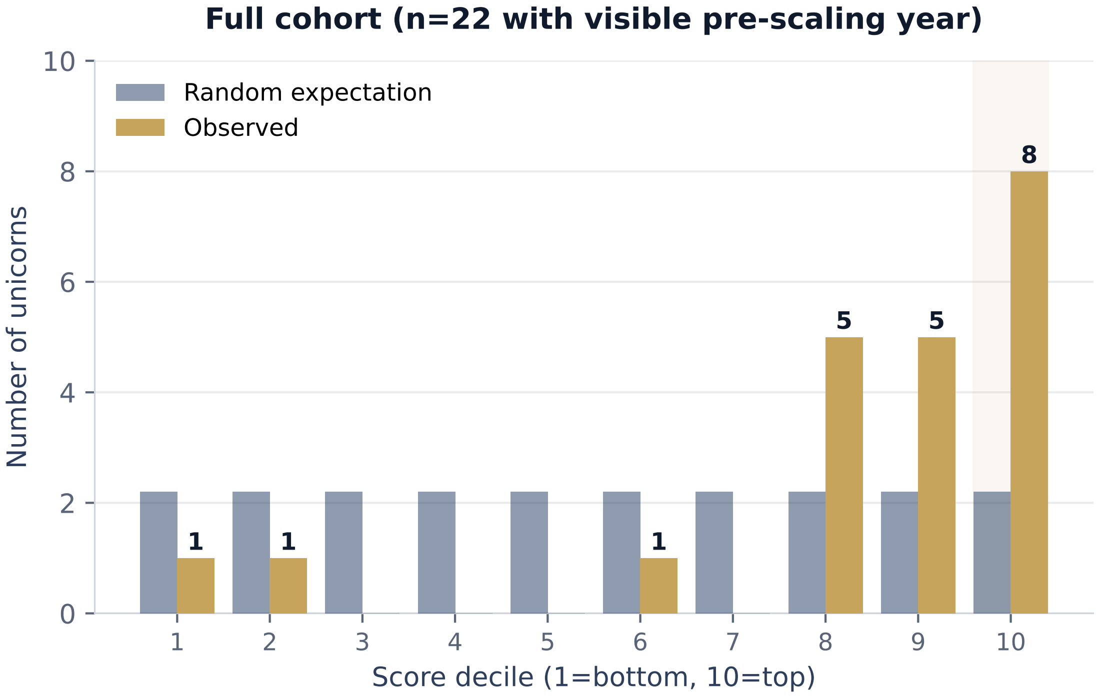
Of those 22, 36% landed in the top decile of model scores for their year, 59% in the top 20%, 82% in the top 30%, and 86% in the top 50%. Random expectation would be 10%, 20%, 30%, and 50% respectively.

Six of the 22 were held out of the training set entirely. The model never saw their data. For this cleanest subset, 50% landed in the top decile and 100% landed in the top 30%.

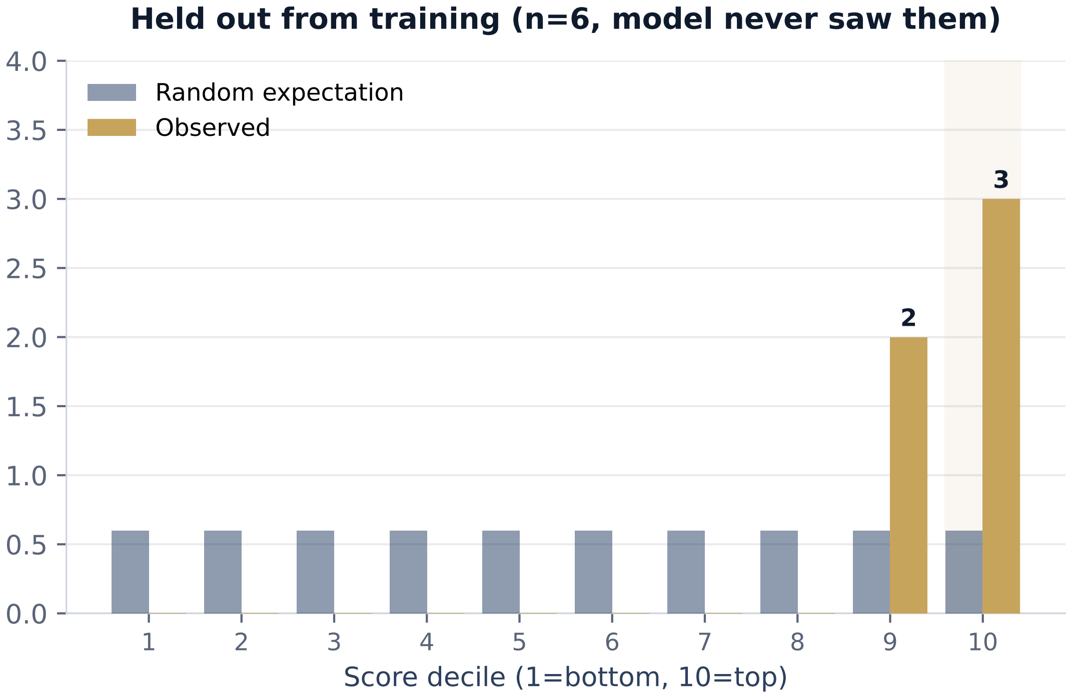
Companies that subsequently became unicorns showed, at the year before their measurable scaling event, workforce signatures the model recognized as pre-hypergrowth.

Unicorns concentrate in top scores

Held-out unicorns (right) provide the cleanest validation. The model placed 100% of them in the top 30% at their pre-scaling year.



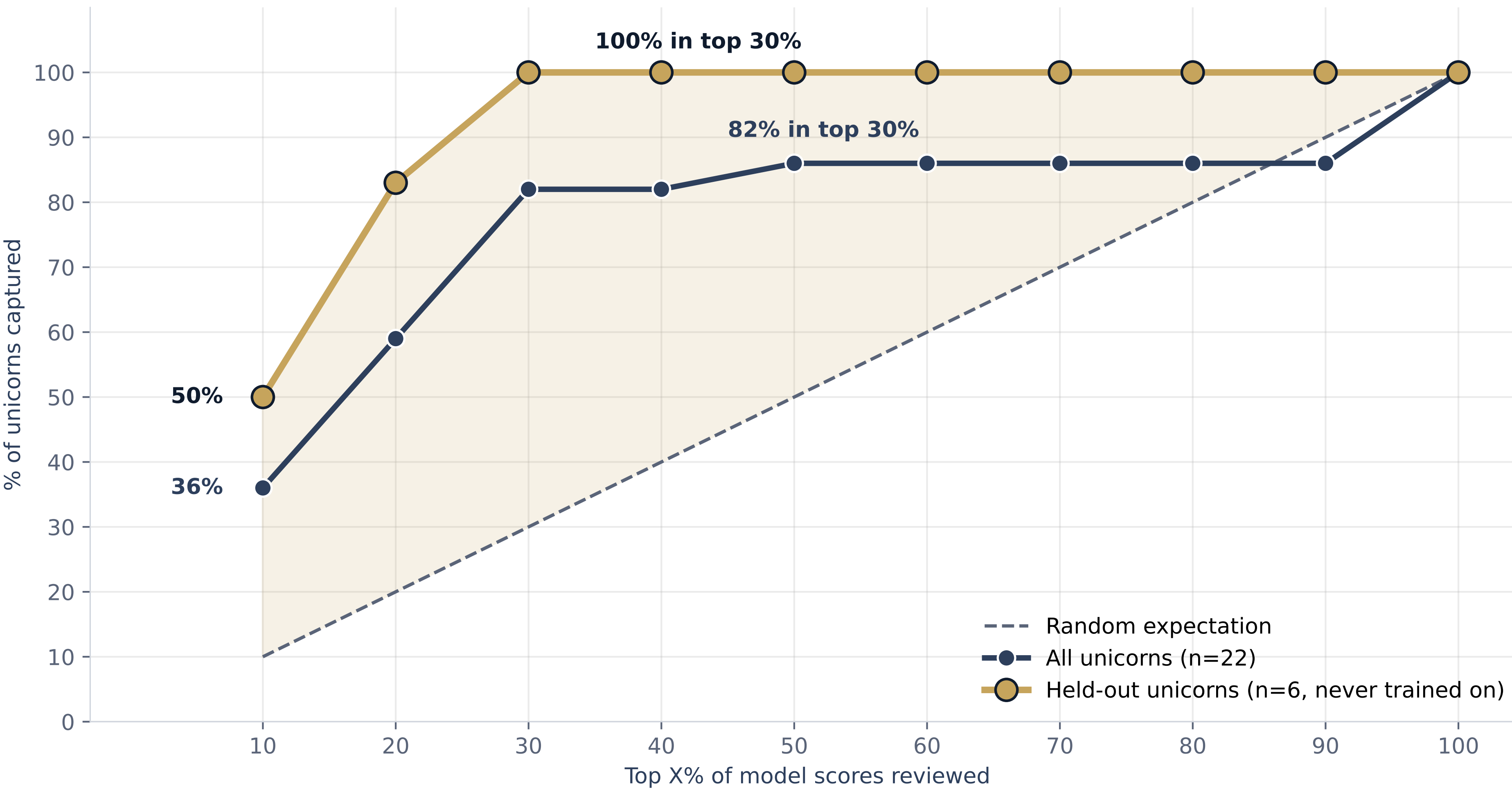
Top decile: 36% (3.6x) · Top 30%: 82% (2.7x)



Top decile: 50% (5.0x) · Top 30%: 100% (3.3x)

Operational view: capture rate by review depth

Reviewing the top 30% of model scores captures 100% of held-out unicorns and 82% of all unicorns at their pre-scaling year.



Source: Vivameda Workforce Panel · 43,208 companies, time-safe features, company-level train/test split

From whitepaper to AI agent

The model in this paper is not only a score. The same workforce panel, the same signal definitions, and the same model logic are packaged into a knowledge file that an AI agent can ingest directly. The buyer does not need to rebuild 70 years of workforce history. Vivameda supplies the pre-computed signal layer and the interpretation logic.

Once ingested, an AI agent can answer questions about any company in your pipeline, portfolio, or research universe:

- Which companies in my pipeline resemble historical pre-hypergrowth companies?
- Which accounts show structural scaling readiness?
- Which companies have momentum but weak workforce structure?
- Which companies have organizational maturity before the growth curve becomes obvious?
- Why is this company scoring high or low? What features drive the score?

The knowledge file works in ChatGPT, Claude, custom agent frameworks, and standard RAG infrastructure (Pinecone, Weaviate, Chroma, and others). It contains the signal definitions, the score interpretation logic, and pre-computed signals on the validation cohort, with panel-scale signal layers available under license. The agent reasons over the file as text, the same way it reasons over any structured knowledge source.

DOWNLOAD THE SAMPLE

Ingest into ChatGPT or Claude in 5 minutes.

Upload the markdown file as a knowledge source. The assistant activates as a Unicorn Signature Analyst and is ready to evaluate companies against the 22 historical pre-scaling signatures.

vivameda.com/Vivameda_Unicorn_Signature_RAG_Sample.md

What this means

The model produces a score from observable company data. The score concentrates a real predictive signal at the top of the distribution. Three audiences should read this differently.

For investors.

Funnel compression and attention allocation. A reviewer screening only the top 30% of model scores would have captured 82% of unicorns with visible pre-scaling years, and 100% of the held-out unicorns. Random review of the same 30% would capture 30%. The model does not guarantee winners; it concentrates probability. For investors facing thousands of inbound companies, the compression is the value.

For accelerators and platforms.

Application screening, founder benchmarking, and program selection. Companies in the top 1% of scores enter hypergrowth at 27x the base rate. Top 1% precision is 11%, so most top-ranked companies still do not hypergrow. The signal is sharp at the top of the distribution and progressively softer below; the right operational mode is ranked review, not threshold rejection.

For AI agents and GTM platforms.

Account scoring, research automation, and explainable company intelligence. The model logic and pre-computed signals are available as a knowledge file, allowing AI agents to reason about scaling readiness for any company in your CRM, portfolio, or research universe. Growth tells you what is moving; the workforce signals explain whether the organization has the shape of a scaler.

Caveats and limitations

Hypergrowth is the prediction target, not unicorn status.

The model predicts hypergrowth, defined as >100% year-over-year workforce growth within a 3-year window. Hypergrowth is a precursor to unicorn outcomes but is not identical to unicorn status. The model is not a direct unicorn classifier.

Workforce features overlap with growth in the full model.

The strongest baseline (growth, industry, age) slightly outperforms the full model on AUC. Workforce features still show standalone predictive signal (workforce-only AUC 0.658, top 1% lift 4.4x) and provide interpretability and enrichment value for AI agents reasoning about scaling readiness. The commercial value of the workforce layer is as enrichment and explanation, not as replacement for growth-based scoring.

Sample sizes for the unicorn validation are modest.

Of 50 named tech unicorns, 22 had a visible pre-scaling year and 6 were held entirely out of training. The held-out sample produces the cleanest measurement and the direction is clear, though estimates from a 6-company subset carry meaningful uncertainty.

Selection effects and survivor bias.

The 50 named unicorns are selected for reaching unicorn status before 2024 and having sufficient data in our panel. This biases the cohort toward US tech-adjacent companies. Concentration of unicorns in top deciles demonstrates that unicorns shared the predictive signature, not that all companies with the signature became unicorns.

Workforce coverage.

Our panel observes workforce composition through public sources covering 4 million companies across 70 years. Trajectories, growth rates, and pattern firings are reliable as relative measures across companies and time periods.

Appendix

The 50-unicorn validation cohort

- Airbnb
- Asana
- Atlassian
- Box
- Coinbase
- Confluent
- CrowdStrike
- Databricks
- Datadog
- DataRobot
- DocuSign
- DoorDash
- Dropbox
- Elastic
- Epic Games
- Fastly
- Flexport
- GitHub
- HashiCorp
- HubSpot
- Instacart
- Jamf
- Klarna
- Lyft
- MongoDB
- MuleSoft
- Opendoor
- PagerDuty
- Pinterest
- Redfin
- Rivian
- Robinhood
- Roblox
- Salesforce
- Samsara
- Segment
- ServiceNow
- Shopify
- Slack
- Spotify
- Splunk
- Square
- Stripe
- Tanium
- Uber
- Unity
- WeWork
- Workday
- Zillow
- Zscaler

Test set summary

Test set: 90,233 company-year observations across 12,285 companies, none of which appear in the training set. 366 positive hypergrowth outcomes (0.41% base rate). Top 1% of test set contains 902 observations and 100 positive outcomes, meaning 1% of reviewed observations captures approximately 27% of all positives in the held-out test set.

Model details

Gradient boosted tree classifier, 200 trees, max depth 3, learning rate 0.05. Trained without class balancing. All metrics in this paper are from the held-out test set. Lift numbers in extreme tails carry wide confidence intervals due to small positive counts at the top of the distribution.

What is available

THE UNDERLYING DATASET

4 million companies. 48 million observations. 70 years.

Longitudinal workforce data from 1950 to 2020. The substrate behind every Vivameda product.

The Unicorn Signature is one product built on the Vivameda Longitudinal Workforce Panel. The panel itself is available in four configurations, scaled to your use case.

- **RAG-ready signal file**
  Drop the knowledge file into ChatGPT, Claude, or your agent framework. Your AI reasons about scaling readiness on any company using the validated unicorn signatures as analogical evidence. Suitable for research workflows and AI-agent products.

- **Custom scoring run**
  Send us a company list (your CRM, your portfolio, your research universe). We score it with the Unicorn Signature model and return a labeled feature file with model probabilities, decile placements, pattern flags, and explanation fields. Suitable for production-scale screening on a defined universe.

- **Panel-scale signal license**
  License pre-computed scores and pattern flags across the full Vivameda panel of 4 million companies. Delivered as a labeled feature file. Suitable for ongoing enrichment of your data platform or alternative-data product.

- **Full dataset for machine learning**
  License the underlying longitudinal workforce dataset for training your own models, building proprietary signal layers, or running custom research at 70-year time depth. Suitable for quantitative funds, AI research teams, and academic work.

NEXT STEP

Discuss your use case with us.

Book a 30-minute data discussion call. We will walk through your use case and the right starting point, whether that is the RAG file, a custom scoring run, the panel-scale signal license, or full dataset access.

BOOK A CALL

calendly.com/oli-nold/data-discussion-call

or email oli@vivameda.com