

Vivameda

Data Dictionary

Longitudinal company-evolution dataset. 4.2M companies, 48M observed records, 1950 to 2020. Structured at the company-year grain. **Frozen Point-in-Time Temporal Substrate.** No PII.

4.2M COMPANIES	48M OBSERVED RECORDS	71 YEARS (1950–2020)	226K+ ANALYTICAL CORE	100+ COUNTRIES
--------------------------	--------------------------------	--------------------------------	---------------------------------	--------------------------

OVERVIEW

Why we built this

Most AI systems reasoning about companies suffer from severe recency bias. Their training data is scraped from the last few years of the web, which means they have only ever seen the post-2015 platform economy.

Vivameda fixes that. It is a seventy-year longitudinal panel of 4.2M companies, year by year, from 1950 through 2020. It spans multiple economic regimes: postwar expansion, stagflation, the personal computer era, the dot-com boom and bust, the 2008 financial crisis, the platform era, and the late-cycle zero-interest-rate period. Each row describes a single company's observable state in a single year: its size, growth, workforce structure, capability mix, and computed signal flags.

Vivameda is architected as a **Frozen Point-in-Time Temporal Substrate**. By maintaining a hard historical ceiling at 2020, the asset explicitly eliminates look-ahead bias and temporal data leakage. This static temporal design guarantees absolute cross-run reproducibility for complex hyperparameter tuning and model optimization across distributed GPU clusters. It is engineered from the ground up as a foundational training substrate, not a live feed.

FOR ML TEAMS

- **Static dataset.** Reproducible training runs. The data does not drift.
- **Temporal regime coverage.** Models see more regime variety than they do from any recent-only snapshot.
- **Honest NULL patterns.** NULL cells reflect insufficient signal density at that point in time. NULL is meaningful. Filter accordingly.
- **Aggregation level.** All data is aggregated at the company-year grain. No individual-employee records are present. No PII.
- **Join key.** Use (company_id, year) as the composite primary key.

DATA PROVENANCE

Where the data comes from

Vivameda is built from publicly available workforce information collected by Vivameda over the past decade — including business registries, official disclosures, and public professional and employment data — restructured in-house into the company-year panel format described below.

No personally identifiable information is present. All signals are aggregated at the company-year level. Coverage and density vary by layer: not every column is populated in every row. The record_depth column labels what is available in each row, so consumers can filter cleanly to the level of enrichment they need.

SCHEMA

Column reference

The full dataset follows the schema described below. The same schema applies at universe scale (4.2M companies, 48M records) and in any sample drawn from it.

Identity and metadata

Column	Type	Description
company_id	text	Stable lowercase-dashed identifier used as the join key. URL-safe slug. Primary key for cross-referencing.
company_name	text	Human-readable display name. Title-cased with preserved acronyms (IBM, AT&T, NVIDIA, JPMorgan).
company_domain	text	Primary web domain if known. May be NULL for entities without a standard commercial web presence.
industry	text	Industry classification from a normalized taxonomy (75 categories). Title-cased. See Appendix A.
country	text	Country of corporate headquarters. Title-cased country names. 100+ countries covered.
year	integer	Observation year. Values range 1950 to 2020 inclusive.

Record depth

Column	Type	Description
record_depth	text	Enrichment layers populated in this row. One of: panel_only (base workforce signal only), panel_plus_cap...

Workforce and tenure

Column	Type	Description
headcount_observed	integer	Observed employee count for the company-year. Ranges from 1 to ~180,000+. Directly observable from s...
company_size_midpoint	integer	Midpoint of the size bucket. Useful for numeric aggregation.
company_size_bucket	text	Categorical size bucket. Values: 1. 1-49, 2. 50-199, 3. 200-999, 4. 1000-4999, 5. 5000+.
net_headcount_change	integer	Year-over-year change in headcount. NULL in a company's first observed year.
avg_tenure_years	float	Average employee tenure at the company in that year, measured in years, to 2 decimal places.
tenure_bucket	text	Categorical tenure bucket. Values: 0. Unknown, 1. <2 years, 2. 2-4 years, 3. 4+ years.

Growth

Column	Type	Description
growth_rate_yoy	<i>float</i>	Year-over-year growth rate as a percentage (38.18 means +38.18%). Computed from headcount deltas. N
prev_growth_rate_yoy	<i>float</i>	Prior year's growth_rate_yoy, attached to the current row for convenience in signal computation and patter
growth_is_reliable	<i>bool</i>	Indicates whether the growth rate passed internal plausibility checks. ML trainers may want to filter to grow
growth_bucket	<i>text</i>	Categorical bucket. Values: 0. Unknown, 0. Unreliable, 1. Shrinking fast, 2. Shrinking, 3. Stable, 4. Growing

Role mix (2010 to 2020 only)

The role layer describes the dominant role composition of the company's workforce, sourced from normalized role classifications across a canonical taxonomy of 9 role buckets. See Appendix C.

Column	Type	Description
primary_role_bucket	<i>text</i>	Dominant role category. Values: operations, specialist, engineering, sales, marketing, finance, executive, c
primary_role_pct	<i>float</i>	Percentage of the workforce whose role maps to primary_role_bucket. Range 0-100, 2 decimals.
distinct_role_buckets	<i>integer</i>	Count of distinct role buckets represented in the workforce. Range 1-9.
role_diversity_bucket	<i>text</i>	Categorical diversity label. Values: 2. Medium diversity, 3. High diversity.
role_coverage_pct	<i>float</i>	Percentage of workforce for which role data is successfully mapped. Indicates completeness of the role sig

Capability mix (1990 to 2020)

The capability layer describes what a company's workforce can do, expressed via the canonical skill-bucket taxonomy. For each company-year, the top three capability buckets are ranked by workforce share. The _pct columns give the percentage of the workforce holding skills in that bucket. At universe scale: 1.88B skill rows, 46.5M capability buckets observed. See Appendix B.

Column	Type	Description
top_capability_1	<i>text</i>	Capability bucket with the highest workforce share. NULL for years before 1990.
top_capability_1_pct	<i>float</i>	Share of workforce whose skills map to top_capability_1. Range 0-100, 2 decimals.
top_capability_2	<i>text</i>	Second-ranked capability bucket.
top_capability_2_pct	<i>float</i>	Share for capability rank 2.
top_capability_3	<i>text</i>	Third-ranked capability bucket.
top_capability_3_pct	<i>float</i>	Share for capability rank 3.
capability_coverage_pct	<i>float</i>	Percentage of workforce for which capability data is successfully mapped. Indicates density of the capabili

Signal flags

Four computed binary flags indicating notable year-over-year dynamics. Each flag is 1 (yes), 0 (no), or NULL (not computable, typically the first year of a company's record). All signals are derived solely from growth_rate_yoy, prev_growth_rate_yoy, and headcount_observed. Definitions are deterministic and reproducible.

Column	Type	Description
early_scaling_flag	int/NULL	1 when growth_rate_yoy > 30 AND headcount_observed BETWEEN 50 AND 500 AND prev_growth_rate_yoy > 0. Captures early scaling in small companies.
contraction_flag	int/NULL	1 when growth_rate_yoy < -5 AND prev_growth_rate_yoy > 0. Captures transitions from growth to meaningfully negative growth.
recovery_signal_flag	int/NULL	1 when growth_rate_yoy > 5 AND prev_growth_rate_yoy < -5. Captures reversals from decline to growth.
growth_acceleration_flag	int/NULL	1 when growth_rate_yoy > prev_growth_rate_yoy AND growth_rate_yoy > 10. Captures increasing growth rates.

On signal definitions. These definitions are documented so any downstream consumer can reproduce them exactly. Many ML trainers will prefer to compute their own signals from the raw growth fields, tailored to specific use cases. The raw fields (growth_rate_yoy, prev_growth_rate_yoy, headcount_observed) are always available in the data, including rows where the flags themselves are NULL.

COVERAGE

Layer coverage by era

Not every row has every layer populated. Coverage layers onto the base panel as source density increases through time. This is the universe-level structure of the full 48M-record dataset.

Era	Record depth	Base panel	Capability	Role	Signals
1950–1989	panel_only	yes (limited)	—	—	computed where possible
1990–2009	panel_plus_capability	yes	yes	—	yes
2010–2016	panel_plus_capability_plus_role	yes	yes	yes	yes
2017–2020	full_record	yes	yes	yes	yes

Pre-1990 rows contain only base panel data (headcount, growth, tenure). Capability and role data did not exist at the source density required for credible enrichment before these dates. Signal flags are computable whenever `prev_growth_rate_yoy` is available, typically from a company's second observed year onward. First-year-of-company rows have NULL signal flags by design.

Analytical core vs. full universe

Of the 4.2M companies in the universe, **226K+ form the high-density analytical core**. These are companies with sustained multi-layer coverage and sufficient observation density for advanced longitudinal signal. The broader universe is useful for breadth (industry coverage, geographic diversity, long-tail exposure). The analytical core is where training-grade signal is deepest.

Underlying the company-year panel: **1.88B skill rows and 46.5M capability buckets** power the capability and role layers. These are the raw source densities from which the company-year aggregates are computed.

SAMPLE

The 503-company evaluation sample

A curated 503-company sample is available for direct evaluation. It is drawn from the universe and follows the same schema documented above. It exists to let prospective consumers inspect the structure and behavior of the data at a scale small enough to work with directly.

503 COMPANIES	25,988 COMPANY-YEAR RECORDS	1950–2020 YEARS COVERED
-------------------------	---------------------------------------	-----------------------------------

Sample filename: vivameda_longitudinal_sample_503companies_1950_2020.csv

Sample tier composition

The 503 anchor companies are curated across five historical tiers, each representing a different depth band. This curation ensures the sample surfaces companies with deep enough observation density to exercise every layer of the schema.

Tier	Companies	Description
70yr_anchor	99	Observable across the full 1950 to 2020 range. Large
60-69yr_anchor	147	Observable for 60–69 years. Postwar-era giants and l
50-59yr_anchor	147	Observable for 50–59 years. Late-20th-century emerg
40-49yr_anchor	99	Observable for 40–49 years. Tech and consumer com
modern_showcase	11	Young high-profile companies founded 1990s–2010s:

Sample coverage by era

The 25,988 sample rows break down across the four record_depth layers as follows:

Era	Record depth	Rows
1950–1989	panel_only	11,182
1990–2009	panel_plus_capability	9,318
2010–2016	panel_plus_capability_plus_role	3,476
2017–2020	full_record	2,012

The sample includes the sample_tier column to indicate which tier each company was drawn from. This column does not exist at universe scale. It is a sample-specific metadata field to help evaluators reason about the curation.

APPENDIX A

Industry taxonomy

The dataset uses a normalized industry taxonomy of 75 categories. Below is the complete list with row counts as observed in the 503-company sample. At universe scale all 75 categories are populated with substantially deeper coverage.

Industry	Rows	Industry	Rows	Industry	Rows
Higher Education	3,029	Medical Devices	229	Biotechnology	108
Retail	2,775	Entertainment	215	Consumer Electronics	108
Financial Services	2,217	Staffing and Recruiting	213	Facilities Services	107
Hospital & Health Care	1,867	Non-Profit Org. Mgmt	204	Apparel & Fashion	102
Information Technology and Services	1,660	Accounting	197	Wholesale	94
Insurance	1,012	Cosmetics	191	Environmental Services	78
Military	704	Consumer Goods	184	Printing	71
Pharmaceuticals	688	Health, Wellness and Fitness	180	Law Enforcement	71
Government Administration	619	Hospitality	175	International Affairs	71
Telecommunications	578	Semiconductors	164	Business Supplies & Equipment	70
Restaurants	529	Human Resources	160	Paper & Forest Products	70
Oil & Energy	507	Construction	160	Farming	70
Banking	465	Civil Engineering	151	Information Services	70
Food & Beverages	401	Broadcast Media	148	Mech./Industrial Engineering	66
Automotive	371	Research	148	Capital Markets	62
Real Estate	360	Machinery	141	Sporting Goods	57
Airlines/Aviation	329	Utilities	139	Wine And Spirits	53
Defense & Space	321	Package/Freight Delivery	136	Media Production	51
Education Management	312	Logistics and Supply Chain	134	Leisure, Travel & Tourism	43
Internet	295	Investment Banking	129	Supermarkets	41
Computer Software	290	Food Production	126	Computer Networking	37
Aviation & Aerospace	286	Design	113	Wireless	36
Chemicals	284	Security and Investigations	112	Computer Hardware	28
Management Consulting	266	Transportation/Trucking/Rail	112	Outsourcing/Offshoring	23
Electrical/Electronic Mfg	259	Commercial Real Estate	110	Packaging and Containers	6

APPENDIX B

Capability taxonomy

The `top_capability_*` columns draw from a canonical taxonomy of skill buckets. The 14 buckets below are those observed as top-ranked in the 503-company sample. The universe-scale taxonomy covers a broader set, filtered down to these 14 as the dominant top-ranked categories for the sample's 1990 to 2020 capability-eligible rows.

communication	customer_service	data_analytics	design
education_training	finance_accounting	healthcare	hr_recruiting
leadership	office_productivity	project_management	research
sales_marketing	software_engineering		

APPENDIX C

Role taxonomy

The `primary_role_bucket` column draws from 9 canonical role buckets:

operations	specialist	engineering
sales	marketing	finance
executive	customer	hr

Next step

Book a 20-minute call to discuss how Vivameda fits your training stack.

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