Decision-Conditioned Company Intelligence

A Multi-Lens Framework for Structure Discovery, Peer Analysis, and Explainable Risk

Cat A: Data Warriors

Lim Shi Ying, Qian Yunhan, Yang Xiuli

Contents

1. Introduction

This project develops a data-driven company intelligence prototype designed to support informed decision-making using large-scale firm-level data. Instead of analysing companies as standalone entities, the system adopts a comparative, peer-based perspective, grouping firms according to shared operational, organisational, and firmographic characteristics. This peer-based framing enables users to interpret company attributes in context, rather than relying solely on absolute values or isolated metrics.

The primary objective of the prototype is to bridge the gap between raw company data and actionable business insight. Through systematic data cleaning, feature transformation, and unsupervised learning techniques, the system converts granular inputs such as revenue, employee counts, industry classifications, and ownership structures into higher-level indicators of similarity, scale, and organisational complexity. This enables users to identify meaningful differences in how companies operate, grow, and structure themselves, which are often obscured in traditional descriptive analysis.

2. Dataset Overview

2.1 Basic EDA

The analysis began with an initial exploratory data analysis to assess the structure, completeness, and distributional properties of the CHAMPIONS Group dataset. The dataset comprises 8,559 company records across 74 variables, covering a broad range of firmographic, financial, operational, and organisational attributes. A preliminary review focused on data types and the extent of missingness across variables, providing an early indication of data quality and usability.

Numeric variables including employee counts, revenue, market value, and IT expenditure were examined using histograms to evaluate their distributions. Logarithmic transformations were applied where appropriate to account for pronounced right-skewness and the presence of extreme values. Categorical variables, such as industry classifications, legal status, entity type, and ownership structure, were analysed using frequency distributions to identify dominant categories and sparsely populated fields.

This initial exploration highlighted several important characteristics of the dataset. Many financial and operational variables exhibited substantial sparsity, with a large proportion of zero or missing values. In particular, zero values in revenue, IT budget, and headcount variables were interpreted as likely indicators of non-disclosure rather than true zero activity. These findings informed key preprocessing choices in later stages, including the treatment of zeros as missing values, the use of robust transformations, and the selection of high-signal features for clustering and segmentation. These distributional characteristics motivated the use of log transformations, robust scaling, and peer-relative comparisons in subsequent stages.

2.2 Data Cleaning and Missing Values

Following the initial exploratory analysis, a systematic data cleaning process was undertaken to improve data quality and ensure suitability for multivariate analysis. Variables with extremely high proportions of missing values were carefully reviewed, and features deemed insufficiently informative or unreliable were excluded from further analysis. This step helped reduce noise and dimensionality while retaining variables with meaningful explanatory power.

Special attention was given to numeric fields such as revenue, IT budget, and number of employees, where a large number of zero entries were observed. Based on domain reasoning and distributional patterns identified during EDA, these zero values were treated as missing (NaN), as they likely reflected non-disclosure rather than true zero economic activity. Treating such values as valid zeros would have distorted distance-based methods and principal component extraction. Temporal variables, such as "Year Found," were validated to fall within realistic bounds and used to calculate company age, creating a standardized metric for organizational maturity. Missing values were explicitly flagged across key columns, producing binary indicators that captured the absence of data. These flags not only informed the preprocessing pipelines but also provided valuable insights into data disclosure patterns, highlighting companies with potentially opaque reporting practices.

Categorical variables were normalized and cleaned to ensure consistency, including industry codes (SIC) and ownership types, preparing the dataset for feature engineering and unsupervised learning. Overall, the cleaning process ensured that the dataset was reliable, interpretable, and ready for downstream analysis while preserving signals inherent in missing or anomalous values.

In addition, explicit restricted-data flags were generated for parent, domestic, and global ultimate ownership fields, capturing intentional non-disclosure and serving as governance-related signals in downstream interpretation.

2.3 Data Transformation

Following cleaning, the dataset underwent extensive transformation to enhance analytical robustness. Numeric variables with extreme right-skew, such as revenue, market value, and IT spend were log-transformed using a logarithm function to mitigate the influence of outliers and allow fair comparisons across firms of different scales. Operational device variables, including the number of PCs, laptops, servers, routers, and storage devices, were often reported in ranges. These buckets were converted to midpoints, with open-ended ranges scaled appropriately, producing continuous numeric values for analysis. A DeviceInfoCount metric was created to summarize overall device coverage per company, offering insight into IT footprint and data completeness. To prevent extreme values from disproportionately influencing distance-based methods, numeric variables were winsorized at upper and lower percentile thresholds prior to standardization.

Categorical variables were also transformed to improve interpretability and reduce dimensionality. High-cardinality variables, such as state or province, were top-k encoded to retain the most frequent categories while grouping the remaining categories as "Other." Corporate hierarchy and ownership information were normalized and transformed into flags, including indicators for parent/ultimate company

matches, cross-border relationships, and ownership opacity scores, allowing the system to assess corporate complexity in a structured manner.

3. Methodology

3.1 Decision-Conditioned Analytical Lenses

To capture the multidimensional nature of how firms operate and create value, the analysis adopts a decision-conditioned, multi-lens framework. Rather than assuming a single, universal notion of firm similarity, we explicitly recognise that similarity is context-dependent and driven by the underlying business question. Different decisions require different comparison sets, and firms that are comparable in one respect may be fundamentally different in another.

Accordingly, each analytical lens is deliberately aligned to a specific business question, and each clustering output is interpreted as a question-specific peer group, rather than as an abstract segmentation.

The firmographic lens addresses the question: “Who looks like me structurally?” This lens captures baseline characteristics such as firm size, sectoral positioning, age, and institutional form. It reflects how firms are conventionally described in economic and financial analysis and produces archetype peers, which are the companies that share broadly similar structural and institutional profiles. These archetypes provide a reference point for understanding how firms are positioned within the broader economic landscape.

The technology intensity lens addresses the question: “Who has a similar level of digital and IT maturity?” This lens focuses on IT expenditure, budget allocation, and technology infrastructure proxies. It captures variation in digital readiness and operational sophistication that is often invisible in traditional firmographic indicators, particularly among firms of similar size or sector. The resulting clusters identify transformation peers, which are especially relevant for technology adoption, digital risk assessment, and transformation planning.

The corporate complexity lens addresses the question: “Who has an unusual or complex organisational structure?” Rather than serving purely as a clustering mechanism, this lens emphasises ownership depth, corporate family size, and parent-subsidiary relationships to surface risk flags and structural anomalies. It highlights firms whose organisational arrangements differ markedly from peers, signalling potential governance, regulatory, or coordination risks.

Taken together, these lenses provide complementary, decision-specific views of firm similarity. A firm may be structurally similar to one set of peers, operationally comparable to another, and technologically aligned with yet another. By preserving these distinctions rather than collapsing them into a single segmentation, the framework enables more nuanced, transparent, and actionable company intelligence.

3.2 Feature Selection and Engineering

Feature selection within each lens was guided by three principles: theoretical relevance, comparability across firms, and interpretability in downstream clustering and decision-making. Variables were selected to represent distinct dimensions of firm behaviour while minimising redundancy across lenses.

For the firmographic lens, variables were chosen to capture organisational scale, economic footprint, and institutional context. Numeric indicators such as total and single-site employee counts, revenue, market value, and corporate family size proxy firm magnitude and operational reach. Categorical variables, including 2-digit SIC industry codes, legal status, ownership type, and entity type, situate firms within their sectoral and regulatory environments. The use of 2-digit SIC codes balances industry specificity with statistical robustness. Derived variables such as company age were included to account for lifecycle effects, distinguishing younger, growth-oriented firms from mature incumbents.

For the technology intensity lens, variables were selected to reflect firms' tangible IT infrastructure and digital investment. Counts of desktops, laptops, servers, routers, and storage devices were included as direct measures of operational IT capacity. As these variables were originally reported in ranges, midpoints were used to enable continuous analysis while preserving relative ordering. To ensure comparability across firms of different sizes, absolute measures were complemented with normalised indicators such as IT spend per employee and devices per employee, allowing the analysis to capture technology intensity rather than scale alone.

For the corporate complexity lens, variables were selected to measure ownership structure, governance depth, and transparency. Indicators capturing the number of corporate family members, parent and ultimate ownership relationships, and domestic versus cross-border ownership were included to reflect organisational layering and jurisdictional dispersion. A set of engineered binary flags, such as own-parent relationships, cross-border ownership chains, and restricted ownership disclosures were aggregated into an ownership opacity score. This composite measure summarises complexity in a form suitable for clustering while retaining economically meaningful variation.

Across all lenses, feature engineering focused on improving comparability and reducing dimensionality without sacrificing interpretability. Normalised metrics and composite indicators were introduced to ensure that clustering results reflect substantive differences in firm behaviour rather than mechanical effects of size or data structure. To ensure analytical rigor and interpretability, the system explicitly separates structure discovery, interpretation, and explanation. Unsupervised models are used solely to identify peer structure, while all risk signals and narratives are derived downstream through transparent, rule-based logic.

3.3 Lens-Based Unsupervised Analysis

3.3.1 Structure Discovery Phase

Following feature selection and transformation, the analysis proceeds to structure discovery, the first layer of the proposed Company Intelligence Engine. Structure discovery aims to identify latent patterns and peer groupings within each analytical lens using unsupervised learning techniques. Importantly, this stage

is intentionally decoupled from any predefined risk definitions or interpretive rules, ensuring that discovered structures emerge directly from the data rather than from subjective assumptions.

For each analytical lens, a dedicated model matrix was constructed (denoted as X_{firm}, X_{tech}, and X_{comp}). These matrices contain only numeric features, with categorical variables encoded using one-hot encoding and numeric variables standardized using z-score normalization. This ensured comparability across features and suitability for distance-based methods.

To explore and visualise latent structure, dimensionality reduction techniques were optionally applied on a per-lens basis. Principal Component Analysis (PCA) was used as a denoising and interpretive tool, allowing inspection of dominant feature contributions and variance structure. For visualisation purposes, Uniform Manifold Approximation and Projection (UMAP) embeddings were generated to produce two-dimensional representations of each lens, facilitating qualitative assessment of cluster separation and overlap.

Clustering was performed independently within each analytical lens using algorithms chosen to match the structural properties of the feature space. For the firmographic and corporate complexity lenses, KMeans clustering was used as the primary segmentation method. In these lenses, the objective is not fine-grained segmentation but the identification of broad, interpretable peer archetypes, making centroid-based clustering appropriate and desirable. In both cases, silhouette analysis revealed clear local optima at moderate values of k, indicating that these spaces support a small number of stable, interpretable peer groups aligned with broad structural archetypes. In contrast, for the technology intensity lens, silhouette scores increased monotonically with larger values of k, indicating continuous internal differentiation rather than a natural discrete partition. In this setting, forcing a fixed number of clusters via KMeans would risk over-fragmentation and arbitrary boundary selection. HDBSCAN was therefore selected for the technology lens, as it allows clusters of varying density to emerge organically while conservatively labelling firms in sparse regions as noise rather than forcing assignment.

This algorithm selection strategy ensures that clustering reflects genuine data structure rather than imposing uniform assumptions across fundamentally different feature spaces.

3.3.2 Lens Scoring and Relative Positioning

The second layer of the system translates unsupervised structure into interpretable, peer-relative indicators. For each analytical lens, a lens score was computed to summarise a firm's overall position within that lens's feature space.

Given that all features within each lens were standardized, the lens score for firm iii was defined as the mean of its standardized feature values within the lens. This composite z-score provides a simple, transparent measure of relative positioning while avoiding the opacity of more complex weighting schemes. Higher scores indicate stronger presence of the characteristics captured by the lens, relative to dataset peers.

To support intuitive interpretation and cross-company comparison, lens scores were converted into percentile ranks and subsequently discretized into quartiles. Quartile ranks reflect a firm's position relative to the overall population, with the top quartile indicating the highest relative intensity under a given lens. This percentile-based framing ensures that interpretations remain context-sensitive and comparable across lenses with differing feature distributions. Quartiles were used to facilitate interpretation and comparison across lenses with different distributions, while preserving relative ordering.

3.3.3. Cross-Lens Contradiction Detection

While lens-specific scores capture relative positioning within individual dimensions, meaningful insights often arise from discrepancies across lenses. To systematically identify such cases, a contradiction detection mechanism was implemented.

For each firm, absolute differences between quartile ranks across all pairs of lenses were computed. The maximum quartile difference serves as a contradiction score, indicating the degree to which a firm's positioning diverges across dimensions. Firms exhibiting large discrepancies, such as being top-quartile in one lens and bottom-quartile in another, were flagged as exhibiting strong cross-lens contradictions.

In addition to the magnitude of the discrepancy, the specific pair of lenses responsible for the contradiction was recorded, enabling targeted interpretation. This allows the system to distinguish, for example, between firms that are technologically advanced but operationally small, versus firms with complex ownership structures that lack corresponding scale. Such contradictions often signal strategic tension, latent risk, or transitional states that would not be visible within any single lens.

3.4 Explainable Risk Archetypes

Building on lens scores and cross-lens contradictions, the system defines a set of named risk archetypes using transparent, rule-based logic. These archetypes are not used in clustering or structure discovery; instead, they serve as an interpretive layer that translates relative positioning into business-relevant risk signals. Rule-based definitions were chosen to ensure transparency, auditability, and ease of communication to non-technical stakeholders.

Each archetype corresponds to a specific pattern of peer-relative inconsistency. For example, scale imbalance risk captures cases where firms exhibit unusually high revenue relative to workforce size, while IT underinvestment risk flags firms whose technological intensity is low compared to peers of similar scale. Corporate complexity mismatch highlights firms with deep ownership structures that are not accompanied by commensurate operational scale, and data opacity risk identifies firms with extensive missing or restricted disclosures that limit reliable assessment.

By grounding each archetype in explicit percentile- or quartile-based rules, the framework ensures that risk signals are interpretable, auditable, and defensible.

3.5 Validation via Negative Controls

Clustering validation was conducted using lens-appropriate robustness diagnostics rather than a single uniform criterion.

For the technology intensity lens, which was clustered using the density-based HDBSCAN algorithm, validation focused on the presence of coherent density structure and the proportion of observations labelled as noise. A meaningful clustering outcome is characterised by the emergence of multiple clusters with non-trivial membership alongside a reasonable noise fraction, indicating that dense regions exist in the data while isolated firms are conservatively excluded rather than forcibly assigned. The observed distribution of cluster sizes and noise points confirms that the technology feature space exhibits genuine local density structure rather than random variation.

For the firmographic and corporate complexity lenses, which were clustered using KMeans, robustness was assessed through stability analysis across random initialisations. Clustering was repeated across multiple random seeds, and similarity between resulting partitions was quantified using the Adjusted Rand Index (ARI). Consistently high ARI scores indicate that cluster assignments are stable and not sensitive to centroid initialisation, supporting the presence of well-defined, reproducible structure in these lenses.

3.6 Integration with LLM-Based Explanation Layer

The final stage of the pipeline integrates the structured outputs (cluster memberships, lens scores, quartiles, contradictions, and risk archetypes) into a controlled large language model (LLM) explanation layer. The LLM is used strictly for narrative synthesis and does not influence clustering, scoring, or risk detection.

Inputs to the LLM are provided in structured form and include peer-relative metrics, identified contradictions, and triggered risk archetypes along with their underlying rule conditions. This design ensures that generated explanations remain grounded in observed data and preserve transparency, while enabling non-technical users to engage with complex analytical outputs.

4. Results

Results are presented lens by lens, followed by cross-lens contrasts, to illustrate how different notions of similarity produce different but complementary peer groupings.

4.1 Individual Lens PCA & Clustering

4.1.1 Firmographic Lens

The Firmographic lens captures variation in structural and reporting characteristics across firms. Principal Component Analysis was applied to the firmographic feature matrix to examine its intrinsic

dimensionality and dominant sources of variation. The cumulative variance profile *(Fig. 1)* indicates a steep initial rise, with the first three principal components explaining approximately 55% of total variance, and the first eight components exceeding 80%. This suggests that firmographic structure is driven by a small number of dominant latent dimensions rather than by diffuse noise across features.

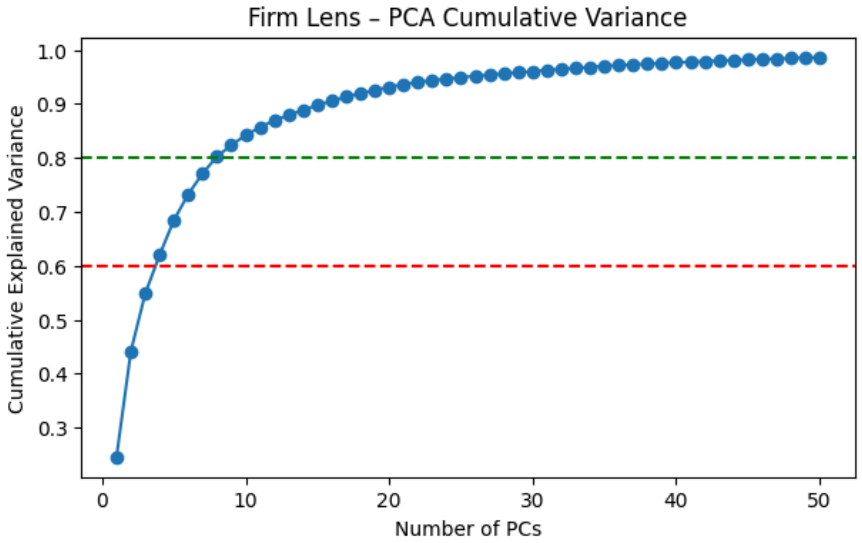

Fig. 1: Firm Lens PCA cumulative variance plot

PCA was used primarily as a diagnostic and interpretive tool rather than as a mandatory dimensionality reduction step, ensuring that clustering results remain grounded in the original feature space.

	PC1	**PC2**	**PC3**
Indicator of	Data completeness + organisational form axis	Economic scale and investment intensity	Lifecycle and organisational role
Top Drivers (Highest Loadings)	• IT spend missing • IT budget missing • Revenue missing • Market value missing • Entity Type (Branch / Subsidiary) • Company age	• IT budget (transformed) • IT spend (transformed) • Revenue • Market value • Employees	• Company age (very high loading) • Employment • Entity type • Corporate family members • Industry (SIC_2D)
Interpretation	Strongly driven by indicators of data completeness and organisational form, separating firms with limited disclosed financial and IT information (such as branches or subsidiaries) from more mature, information-rich entities. This dimension reflects an underlying institutional maturity and reporting opacity axis, rather	Captures firm scale and investment intensity, with high loadings on revenue, market value, employment, and IT expenditure. This dimension aligns closely with conventional notions of economic scale and operational capacity.	Reflects lifecycle and organisational role, distinguishing younger entities from more established firms and separating parents, subsidiaries, and branches with differing employment and structural characteristics.

	than pure scale.		

The two-dimensional projection of firms onto the first two principal components *(Fig. 2)* reveals clear heterogeneity in firmographic structure, with distinct regions corresponding to differences in reporting completeness, organisational form, and scale. While overlap is expected given the continuous nature of firm attributes, the distribution indicates meaningful latent structure suitable for downstream clustering.

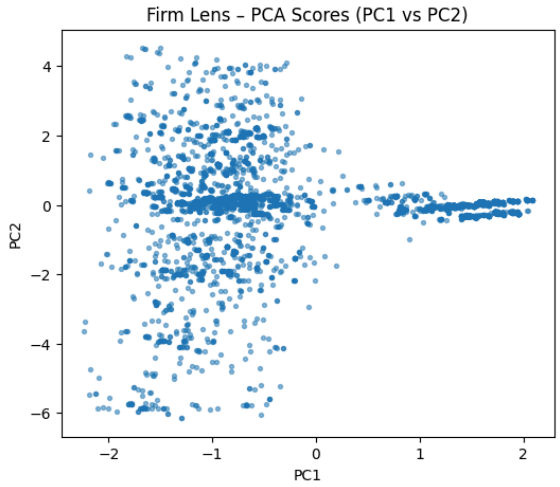

Fig. 2: Firm Lens PC1 vs PC2 scatter plot

Clustering quality was evaluated across a range of cluster counts using silhouette scores. The results *(Fig. 3)* indicate a clear local maximum at k = 3, suggesting that the firmographic space supports a small number of broad, interpretable archetypes. Beyond this point, additional clusters provide limited incremental separation and risk fragmenting economically similar firms. Given the objective of generating stable and interpretable peer groups rather than maximising numerical separation, k = 3 was selected as the baseline firmographic segmentation.

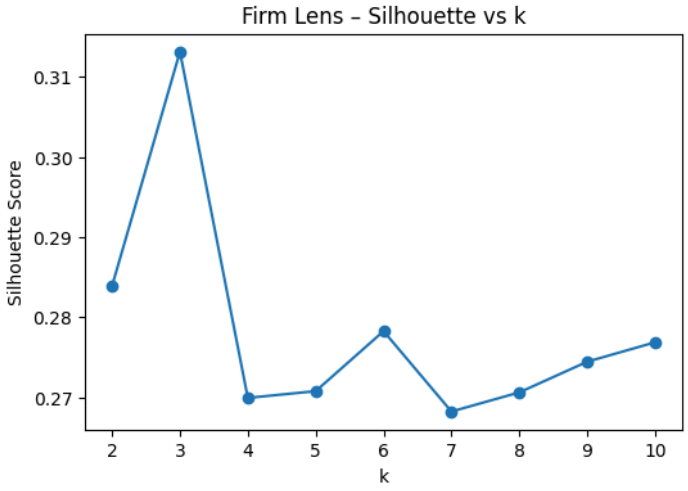

Fig. 3: Firm Lens Silhouette vs k plot

Overall, the firmographic feature space exhibits a low-dimensional structure driven primarily by data completeness, organisational role, and firm scale, supporting the use of a small number of broad peer groups.

4.1.2 Technology Intensity Lens

The cumulative variance profile *(Fig. 4)* indicates a steep initial rise, with the first three principal components explaining approximately 54% of total variance, and the first eight components accounting for nearly 78%. This suggests that variation in firms' technology profiles is driven by a small number of dominant dimensions rather than by diffuse noise across many features. As with the firmographic lens, PCA is used here primarily as a diagnostic and interpretive tool, rather than as a strict dimensionality reduction requirement for clustering. The observed variance structure indicates that the technology lens captures coherent, low-dimensional patterns suitable for peer grouping and downstream interpretation.

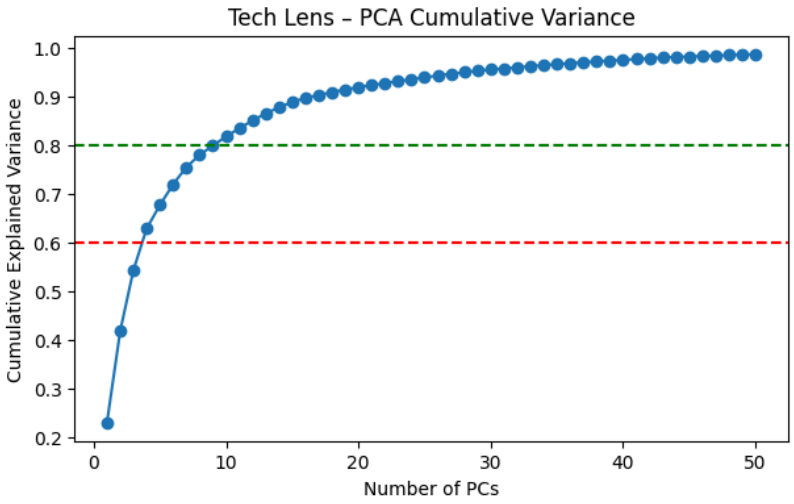

Fig. 4: Tech Lens PCA cumulative variance plot

	PC1	**PC2**	**PC3**
Indicator of	Absolute digital and IT investment intensity	Technology reporting visibility and disclosure completeness	Infrastructure deployment intensity relative to workforce size
Top Drivers (Highest Loadings)	• IT Budget (transformed) • IT Spend (transformed) • IT Spend per Employee (transformed) • IT Spend Missing	• IT Spend Missing • IT Budget Missing • Employees Total (transformed) • Employees Total Missing • Servers per 100 Employees (transformed) • IT Budget (transformed)	• Servers per 100 Employees (transformed) • Employees Total (transformed) • Employees Total Missing • IT Spend per Employee

	• IT Budget Missing • Employees Total Missing • Employees Total (transformed) • Servers per 100 Employees (transformed) • SIC_2D indicators (minor contribution)	• IT Spend (transformed) • SIC_2D indicators (secondary) • IT Spend per Employee (transformed)	(transformed) • IT Spend Missing • IT Budget Missing • Ownership Type (Private / Missing) • SIC_2D indicators (minor)
Interpretation	Dominated by absolute IT investment measures, including IT budget, IT spend, and IT spend per employee. This component captures a firm's overall level of digital and IT investment intensity, separating firms with substantial technology expenditure from those with minimal observed IT investment.	Driven primarily by IT spend and budget missingness indicators, along with employment-related variables. This dimension reflects IT reporting completeness and disclosure visibility, distinguishing firms with transparent technology reporting from those with limited observable IT information.	Shaped by infrastructure density relative to firm size, with strong contributions from servers per employee and IT spend per employee. This component captures infrastructure deployment intensity, distinguishing firms with heavier digital infrastructure from those with lighter technology footprints.

Projection of firms onto the first two principal components *(Fig. 5)* reveals substantial heterogeneity in technology profiles, with firms distributed along continuous gradients of IT investment and reporting completeness. The absence of sharp boundaries is consistent with the expectation that digital maturity evolves gradually, reinforcing the need for clustering and peer-relative comparison rather than rigid classification.

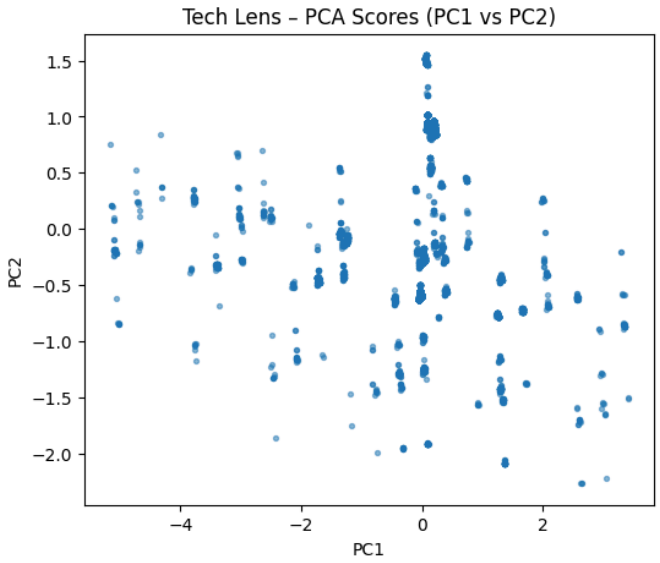

Fig. 5: Tech Lens PC1 vs PC2 scatter plot

Clustering quality was evaluated across a range of cluster counts using silhouette scores. Unlike the firmographic lens, the results *(Fig. 6)* show that silhouette scores increase steadily with larger values of k, indicating that technology intensity exhibits greater internal differentiation and supports finer-grained segmentation. A moderate cluster count was selected to balance interpretability with resolution, yielding technology-based peer groups that reflect meaningful differences in digital investment, infrastructure intensity, and reporting visibility without over-fragmentation.

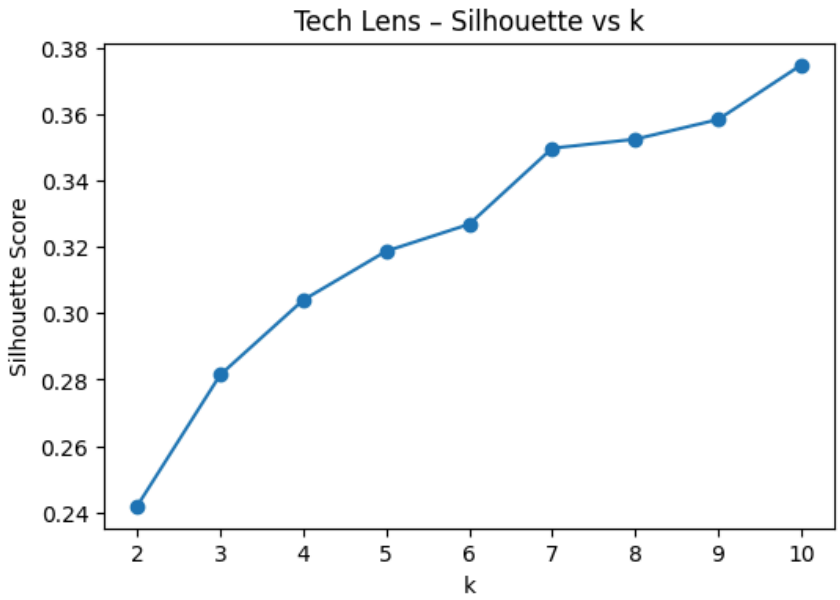

Fig. 6: Tech Lens Silhouette vs k plot

Given the absence of a clear silhouette optimum and the continuous nature of digital maturity, density-based clustering (HDBSCAN) was adopted for the technology lens to avoid arbitrary selection of cluster counts and to allow heterogeneous technology profiles to emerge naturally, with a non-trivial fraction of firms conservatively labelled as noise rather than forcibly assigned, reflecting the underlying density structure of the technology feature space.

4.1.3 Complexity Lens

The cumulative variance profile *(Fig. 7)* exhibits a steep initial rise, with the first three principal components explaining approximately 52% of total variance, and the first eight components accounting for over 76%. This indicates that corporate complexity is governed by a small number of dominant structural dimensions rather than diffuse variation across features.

As with the other lenses, PCA is employed primarily as a diagnostic and interpretive tool. The observed low-dimensional structure suggests that ownership and organisational complexity signals are strongly aligned and well-suited for downstream clustering and peer identification.

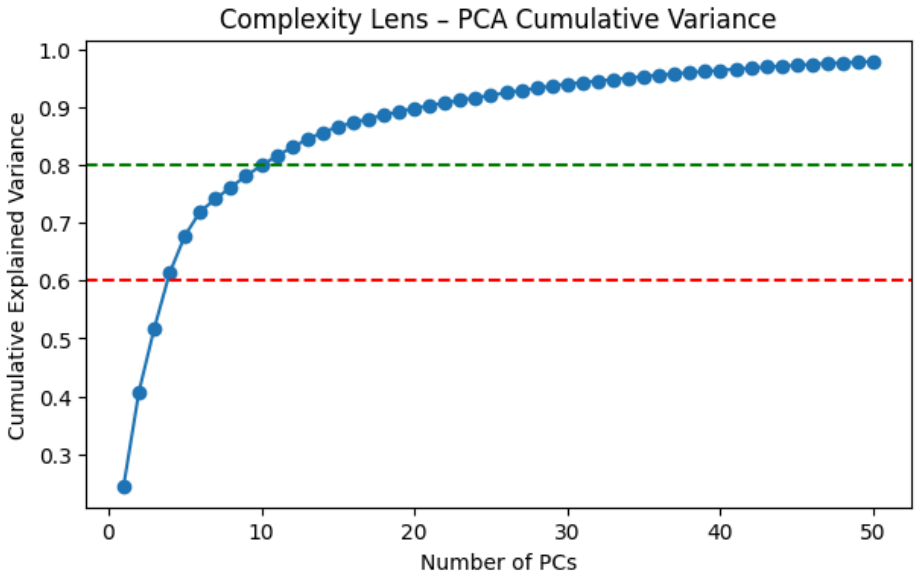

Fig. 7: Complexity Lens PCA cumulative variance plot

	PC1	**PC2**	**PC3**
Indicator of	Ownership Control & Ultimate Status - whether a firm sits at the top of ownership hierarchies or within them	Organisational Layering & Group Position - captures depth within corporate groups and organisational layering	Structural Depth & Opacity - captures ownership complexity combined with information restriction and opacity.
Top Drivers (Highest Loadings)	• IsOwnDomesticUltimate • IsOwnParent • IsOwnGlobalUltimate • Entity Type (Parent) • Corporate Family Members (transformed) • Entity Type (Branch / Subsidiary)	• Entity Type (Subsidiary) • Corporate Family Members (transformed) • Entity Type (Branch) • Ownership Type (Private / Missing) • SIC_2D indicators • Ownership flags (secondary)	• Corporate Family Members (transformed) • Entity Type (Branch / Subsidiary) • Ownership Type (Private / Missing) • Restricted parent indicators • Ownership flags (secondary)
Interpretation	Dominated by ownership structure indicators, including flags for being one's own parent, domestic ultimate, or global ultimate entity, along with entity type and corporate family size. This component captures a firm's position within ownership hierarchies, separating standalone or ultimate entities	Driven primarily by entity role and group depth, with strong contributions from subsidiary and branch indicators, corporate family size, and ownership type. This dimension reflects organisational layering and intra-group positioning, distinguishing firms that	Dominated by corporate family size and restriction indicators, alongside entity type and ownership flags. This component captures structural depth combined with opacity, separating firms with extensive ownership networks and restricted parent information from simpler,

	from firms embedded within multi-layered corporate groups.	operate deep within corporate structures from those closer to the top of ownership chains.	more transparent organisational structures.

Projection of firms onto the first two principal components *(Fig. 8)* reveals clear separation patterns within the corporate complexity space. Compared to the firmographic and technology lenses, the distribution exhibits more distinct structural groupings, reflecting the inherently discrete nature of ownership arrangements and governance roles. This visual structure indicates that firms' positions within corporate hierarchies are a strong and differentiating source of variation.

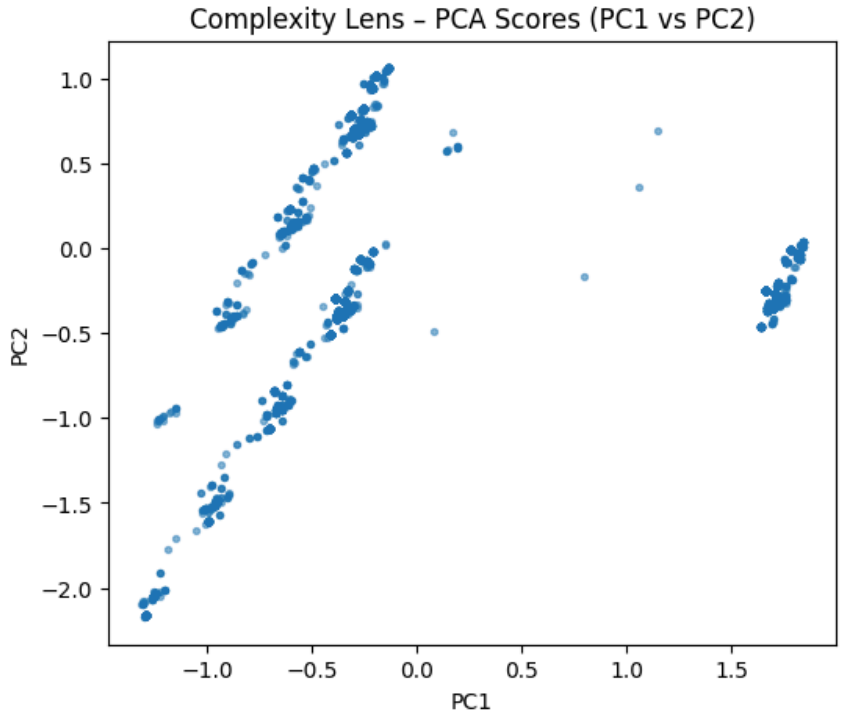

Fig. 8: Complexity Lens PC1 vs PC2 scatter plot

Clustering quality was evaluated across a range of cluster counts using silhouette scores. The results *(Fig. 9)* suggest that corporate complexity lens exhibits consistently higher silhouette scores than the other lenses, with clear local optima at moderate values of k. This indicates that ownership and organisational structure form well-separated and internally coherent clusters, supporting reliable identification of complexity-based peer groups.

A moderate cluster count was selected to balance interpretability with resolution, yielding peer groups that reflect meaningful differences in governance structure, ownership depth, and transparency without excessive fragmentation.

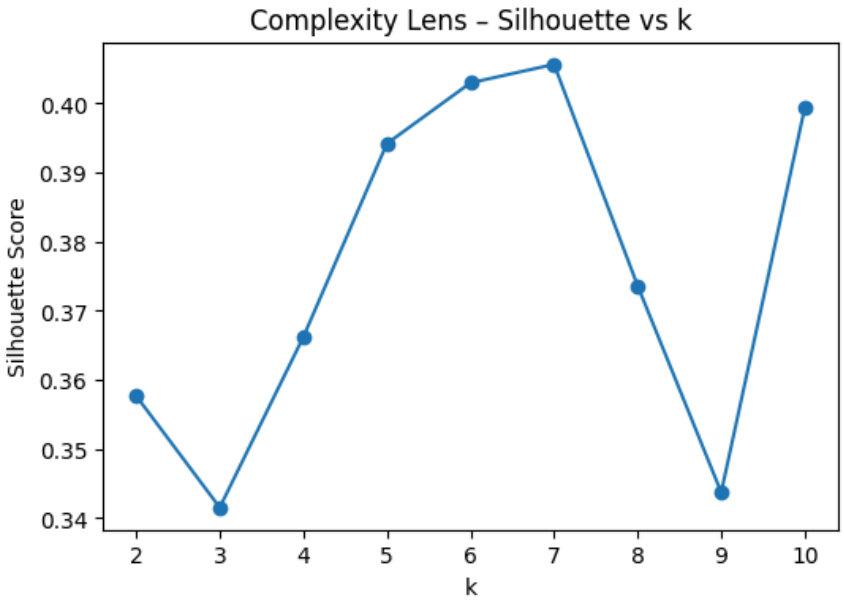

Fig. 9: Complexity Lens Silhouette vs k plot

A sample of corporate complexity lens outputs is presented in *Fig. 10*, illustrating how cluster membership, lens scores, and percentile rankings combine to characterise firms' organisational structures. Each firm is assigned to a complexity cluster based on ownership depth, corporate family size, and governance features. The comp_score reflects a standardized measure of complexity, with positive values indicating above-average structural intricacy relative to peers. To contextualize these scores, percentile ranks are provided both across the full dataset (comp_pct_overall) and within the assigned cluster (comp_pct_within_cluster). These metrics enable identification of both absolute and relative structural differences, supporting nuanced benchmarking, peer comparisons, and the detection of firms with unusual organisational arrangements. Incorporating both overall and within-cluster percentiles ensures that insights capture meaningful variation across the population while remaining sensitive to local peer context.

[comp] lens-score numeric columns: 7

	DUNS Number	Company Sites	comp_cluster	comp_score	comp_pct_overall	comp_pct_within_cluster
0	639677726	Zyf Lopsking Material Technology Co., Ltd. No....	2	-0.202463	0.543580	0.672008
1	547756179	Beijing Kaishi Lide Commerce And Trade Co., Lt...	1	-0.355958	0.109359	0.114619
2	728834216	Keshan Shengren Potato Industry Processing Co....	1	-0.268797	0.420960	0.579602
3	728791839	Zuoquan County Yuanfeng Agriculture Technology...	0	0.756444	0.915294	0.636796
4	728889244	Zuoquan County Tianxin Real Estate Development...	0	0.735618	0.856467	0.309071

Fig. 10: Sample of corporate complexity lens outputs

4.2 Multi-Lens Structure Discovery Outputs

Fig. 11 visualises firmographic clusters projected onto the first two principal components. The clusters occupy distinct regions along the primary axes of firmographic variation, reflecting differences in organisational role, reporting completeness, and scale. While some overlap remains (expected given the continuous nature of firm characteristics), the separation indicates that the clustering captures meaningful structural differences rather than random partitioning.

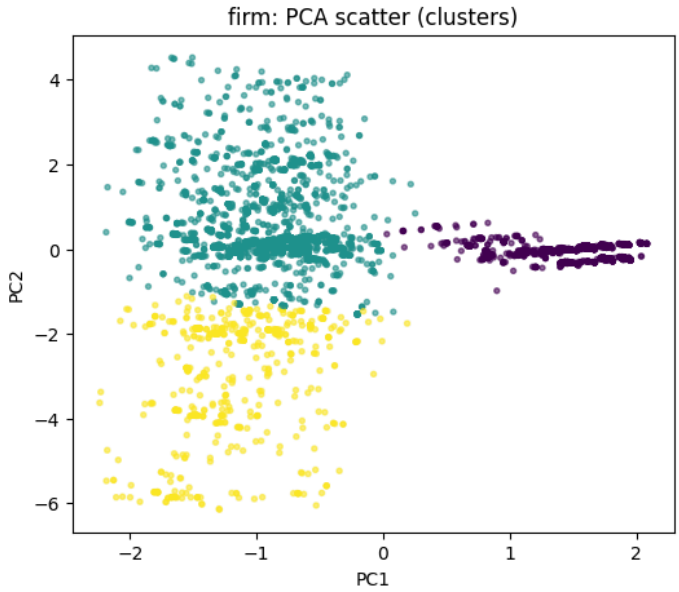

Fig. 11: Firm Lens – PCA Scatter (PC1 vs PC2), coloured by firm_cluster

In the technology lens, density-based clustering (HDBSCAN) identifies multiple technology peer groups while explicitly labelling firms with insufficient local density as noise. As shown in *Fig. 12,* clusters

concentrate in regions of similar IT investment intensity and reporting completeness, while isolated firms are treated conservatively as unclustered rather than forcibly assigned.

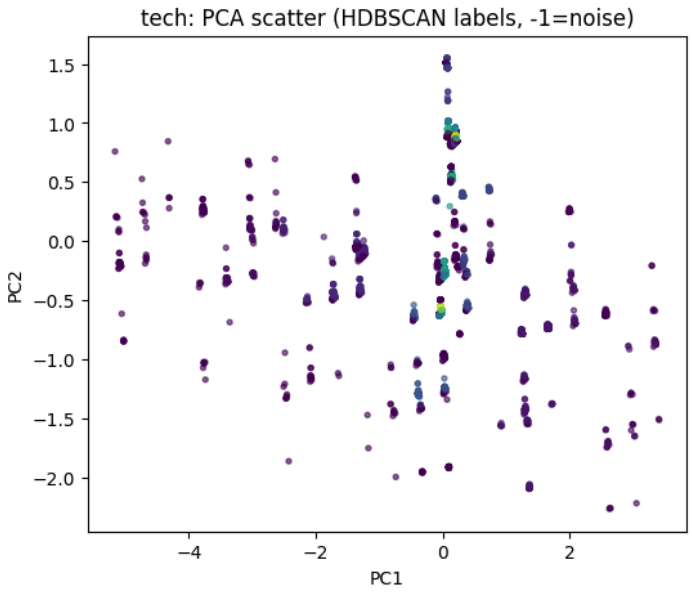

Fig. 12: Tech Lens – PCA Scatter (PC1 vs PC2), coloured by HDBSCAN labels (noise highlighted)

The corporate complexity lens exhibits the clearest visual separation among clusters. As illustrated in *Fig. 13* firms group into well-defined regions corresponding to ownership control, organisational layering, and corporate family depth. This discrete structure is consistent with the inherently categorical nature of governance and ownership arrangements.

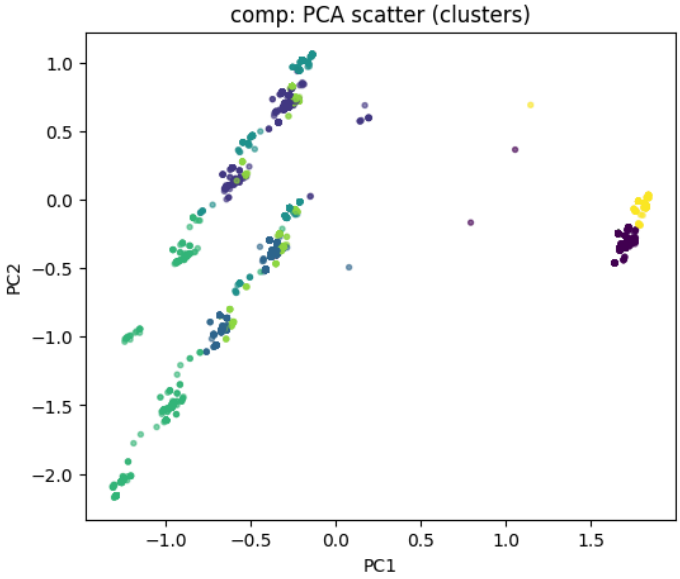

Fig. 13: Complexity Lens – PCA Scatter (PC1 vs PC2), coloured by comp_cluster

For each lens, a composite lens score was computed for every firm as the mean of its standardized numeric features within that lens. These scores provide a continuous summary of a firm's relative position along the dominant dimensions captured by the clustering process.

	firm_pct_overall	q_firm	tech_pct_overall	q_tech	comp_pct_overall	q_comp
0	0.520972	3	0.520154	2	0.543580	3
1	0.155041	1	0.041594	1	0.109359	1
2	0.532539	3	0.175079	1	0.420960	2
3	0.246524	1	0.112805	1	0.915294	4
4	0.123379	1	0.560112	3	0.856467	4

Fig. 14: Example Firm Lens Score Table

To enable peer-relative comparison, lens scores were converted into percentile ranks at two levels: (i) across the full dataset, and (ii) within each lens-specific cluster. Overall percentiles indicate how a firm compares globally, while within-cluster percentiles highlight its position relative to structurally similar peers.

Equivalent score and percentile outputs were generated for the technology and corporate complexity lenses, ensuring consistency across all dimensions of analysis. These outputs form the quantitative basis for subsequent cross-lens comparison, contradiction detection, and risk interpretation.

Taken together, the multi-lens structure discovery process yields three complementary views of firm similarity: archetype peers based on firmographic structure, transformation peers based on technology intensity, and governance peers based on corporate complexity. Each lens produces coherent clusters, interpretable latent dimensions, and peer-relative scores, providing a structured foundation for downstream cross-lens analysis and explainable insight generation.

4.3 Low-Dimensional Visualisation of Multi-Lens Structure (UMAP)

Fig. 15 presents a two-dimensional UMAP projection of the firmographic feature space, coloured by firmographic cluster assignment. The embedding reveals coherent local neighbourhoods corresponding to the identified clusters, with separation primarily along gradients of organisational role, reporting completeness, and firm scale. Some overlap is observed, reflecting the continuous nature of firmographic attributes and reinforcing that cluster boundaries represent soft peer groupings rather than rigid classifications.

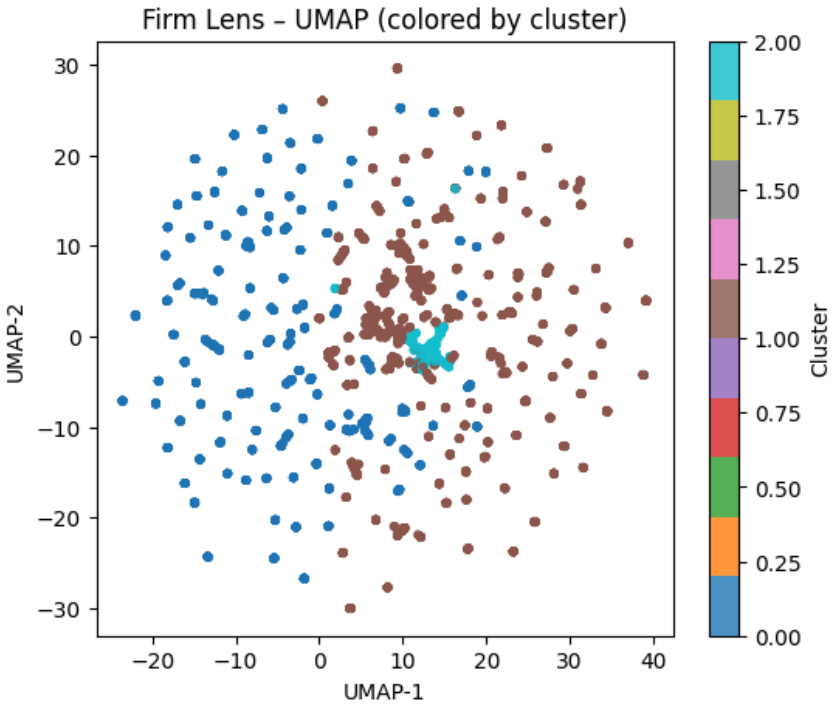

Fig. 15: Firm Lens – UMAP, coloured by firm_cluster

The UMAP embedding for the technology lens *(Fig. 16)* shows a more fragmented structure, with multiple smaller clusters interspersed across the space. This pattern is consistent with the higher internal heterogeneity of technology intensity and digital maturity observed in the PCA and clustering results. Firms identified as noise by the density-based clustering algorithm tend to occupy sparse regions of the embedding, supporting the decision to avoid forced cluster assignment for these observations.

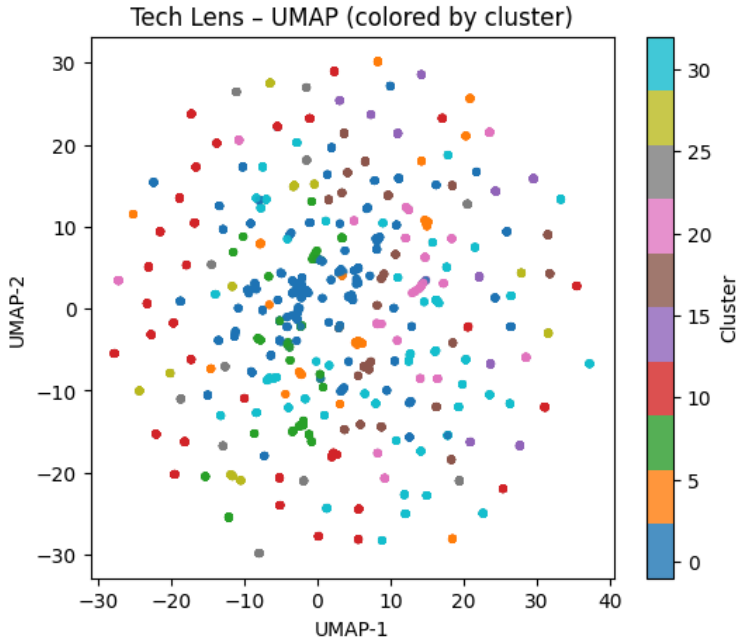

Fig. 16: Tech Lens – UMAP, coloured by tech_cluster (HDBSCAN)

In contrast, the corporate complexity lens exhibits clearer separation in the UMAP embedding *(Fig. 17)*, with clusters forming more compact and distinct regions. This reflects the inherently discrete nature of ownership roles, governance structures, and corporate family configurations. The observed separation aligns with the higher silhouette scores obtained for this lens, indicating well-defined complexity-based peer groups.

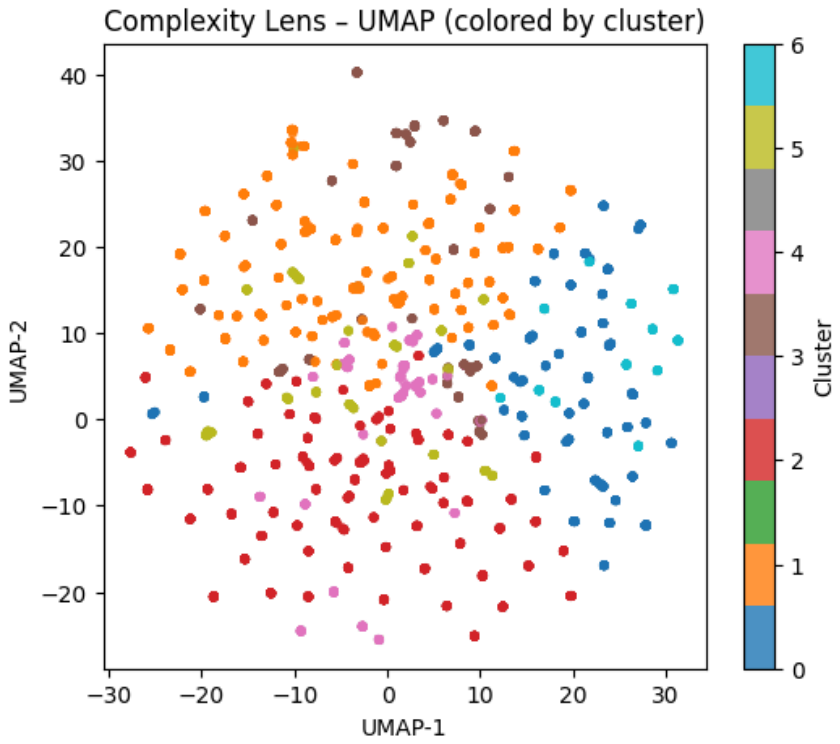

Fig. 17: Complexity Lens – UMAP, coloured by comp_cluster

Across all three lenses, UMAP visualisations provide complementary confirmation of the clustering results. While firmographic and technology structures exhibit gradual transitions and overlap, corporate complexity displays more discrete separation. These differences reinforce the central premise of the framework: firm similarity is lens-dependent, and different business questions surface different structural patterns.

4.4 Cross-Lens Contradiction Analysis

To identify multi-dimensional inconsistencies in firm profiles, contradiction scores were computed as the maximum absolute difference in quartile ranks across the three analytical lenses. Firms exhibiting substantial discrepancies are flagged as having cross-lens contradictions, indicating potential strategic tension, latent risk, or transitional states that are not visible within any single lens.

Fig. 18 presents the contradiction results derived from the lens scores. For each firm, quartile ranks are reported across the three lenses, along with the maximum quartile difference and the specific pair of lenses responsible for the discrepancy.

contradiction_pair	contradiction_strength	contradiction_level	contradiction_quartile_maxdiff
tech vs complexity	0.023426	none	1
firm vs tech	0.113448	none	0
firm vs tech	0.357460	moderate	2
tech vs complexity	0.802489	strong	3
firm vs complexity	0.733088	strong	3

Fig. 18: Cross-Lens Contradiction Results

These results highlight firms that are structurally similar to one set of peers but differ markedly in other dimensions. Such discrepancies may signal strategic misalignment, operational or digital gaps, or potential governance concerns. For example, a firm with high structural scale but low technology intensity may face challenges in digital transformation, whereas a firm with high corporate complexity but modest operational scale may encounter governance or coordination risks.

4.5 Explainable Risk Archetypes

Building on lens scores and cross-lens contradictions, a set of rule-based risk archetypes was defined to provide interpretable, actionable insights. These archetypes serve as an interpretive layer, translating quantitative discrepancies into business-relevant risk signals.

	DUNS Number	Company Sites	risk_tags
0	639677726	Zyf Lopsking Material Technology Co., Ltd. No....	[]
1	547756179	Beijing Kaishi Lide Commerce And Trade Co., Lt...	[]
2	728834216	Keshan Shengren Potato Industry Processing Co....	[IT Underinvestment Risk]
3	728791839	Zuoquan County Yuanfeng Agriculture Technology...	[Complexity Mismatch Risk, Data Opacity Risk]
4	728889244	Zuoquan County Tianxin Real Estate Development...	[Complexity Mismatch Risk, Data Opacity Risk]

Fig. 19: Explainable Risk Archetypes Results

These archetypes contextualise the observed contradictions. For instance, Scale Imbalance Risk flags firms where revenue or market footprint is disproportionately high relative to workforce size, whereas IT Underinvestment Risk identifies firms whose technology intensity is low relative to structural peers. Corporate Complexity Mismatch highlights firms with intricate ownership structures not aligned with operational scale, and Data Opacity Risk identifies firms with missing or restricted data that limit reliable assessment. Grounding these archetypes in explicit, percentile-based rules ensures that risk signals are interpretable, auditable, and defensible, supporting targeted decision-making and strategic review.

5. Insights and Implications

5.1 Firm Similarity Is Lens-Dependent, Not Absolute

The multi-lens analysis demonstrates that firm similarity is inherently context-dependent. Firms that appear comparable under one analytical lens may differ substantially under others, reflecting the multidimensional nature of how organisations operate and create value. The firmographic lens captures structural archetypes based on size, maturity, and organisational role; the technology lens surfaces variation in digital investment and infrastructure intensity; and the corporate complexity lens isolates governance structure, ownership depth, and transparency.

Across lenses, the underlying data exhibit distinct geometric and clustering properties. Firmographic and technology features form largely continuous spaces with gradual transitions, while corporate complexity displays more discrete structural groupings. This divergence underscores the limitations of single-view segmentation approaches and motivates the use of multiple, decision-conditioned lenses to support our nuanced analysis.

5.2 Peer Groups Vary by Business Question

Because each lens encodes a different notion of similarity, the resulting peer groups answer different business questions. Firmographic clusters identify archetype peers (firms that are structurally comparable in scale, age, and organisational form). Technology-based clusters reveal transformation peers, highlighting firms with similar levels of digital investment and IT maturity regardless of size or industry. Corporate complexity clusters define governance peers, grouping firms with similar ownership hierarchies and organisational depth.

Benchmarking a firm’s IT spend against firmographic peers answers a fundamentally different question than benchmarking against technology peers. By preserving these parallel peer definitions, the framework avoids conflating structurally similar firms with operationally or technologically similar ones.

5.3 Lens Scores Enable Peer-Relative Positioning

The introduction of lens-specific composite scores and percentile rankings gives us a consistent, interpretable way to position firms within each analytical dimension. Overall percentiles indicate how a firm compares globally, while within-cluster percentiles contextualise performance relative to structurally similar peers.

These dual perspectives reveal patterns that would be obscured by absolute metrics alone. For example, a firm may appear average in overall technology intensity but rank highly within its firmographic cluster, indicating above-peer digital maturity given its size and structure. Conversely, firms with strong absolute metrics but low within-cluster percentiles may underperform relative to their closest peers.

5.4 Cross-Lens Mismatches Surface Strategic Tensions

Synthesising results across lenses highlights systematic mismatches that carry strategic and risk-related implications. Firms with high firmographic scale but low technology intensity may signal potential IT underinvestment, while highly complex ownership structures combined with limited operational scale may indicate governance or coordination risk. Conversely, smaller firms with disproportionately high technology scores may represent digitally native or transformation-led business models.

These mismatches are not anomalies in isolation; they emerge only through cross-lens comparison. By design, our framework does not force a single interpretation of such divergences, but instead flags them as areas warranting closer examination.

5.5 Structural Patterns Are Robust, Not Artefacts

The consistency of findings across PCA diagnostics, clustering metrics, and UMAP visualisations supports the robustness of the discovered structures. Differences in silhouette scores and visual separability across lenses align with theoretical expectations: ownership structures form more discrete categories than firm size or technology adoption, which evolve more continuously.

Importantly, the interpretation of clusters and scores is grounded in stable, low-dimensional structure rather than in overfitted or purely algorithmic artefacts. This robustness underpins the reliability of downstream insights and supports the use of the framework for exploratory analysis, benchmarking, and risk screening.

5.6 Implications for Decision Support and Data Value

Taken together, the multi-lens framework transforms raw firm-level attributes into actionable company intelligence. It enables users to identify appropriate peer groups for benchmarking, surface structural and operational mismatches, and contextualise firm behaviour across multiple dimensions without sacrificing interpretability.

From a commercial perspective, our approach demonstrates how a single dataset can support diverse analytical use cases, ranging from market segmentation and competitive benchmarking to governance assessment and digital readiness evaluation, by re-framing similarity around the specific decision at hand. This flexibility significantly enhances the value proposition of the underlying data.

6. Conclusion

Our project demonstrates that firm similarity is not an absolute concept, but a decision-conditioned one. Rather than imposing a single, monolithic segmentation, we show that meaningful company intelligence emerges only when similarity is defined relative to the business question being asked. By decomposing firm behaviour into firmographic structure, technology intensity, and corporate complexity, the framework reveals parallel but distinct peer groupings that would be obscured under traditional one-size-fits-all clustering approaches.

Methodologically, the system separates structure discovery from interpretation, ensuring that unsupervised learning is not contaminated by handcrafted risk assumptions. Each lens is allowed to express its own intrinsic geometry: broad archetypes in firmographic and complexity spaces, and heterogeneous, density-driven groupings in the technology space. Validation strategies are explicitly matched to these properties (using clustering stability for centroid-based lenses and density structure diagnostics for the technology lens) demonstrating that the uncovered patterns reflect genuine business structure rather than algorithmic artefacts.

Crucially, insight is generated not from clusters alone, but from relative positioning across lenses. Lens scores, percentile rankings, and cross-lens contradiction detection transform raw clustering outputs into interpretable signals that surface strategic tensions, transitional states, and latent risks. The introduction of rule-based risk archetypes further translates these signals into defensible, auditable indicators that can be acted upon without sacrificing transparency.

Beyond its analytical contributions, the framework illustrates how firm-level data can be elevated from static descriptors into a flexible decision-support system. The same underlying dataset supports benchmarking, transformation assessment, governance screening, and risk identification, simply by changing the lens through which similarity is defined. The controlled integration of an LLM explanation layer reinforces this value proposition, enabling complex, multi-dimensional outputs to be communicated clearly to non-technical users while preserving traceability to underlying data and rules.

In sum, our project moves beyond clustering as an end in itself. It offers a generalisable, explainable, and business-aligned approach to company intelligence, one that respects the multidimensional nature of firms, adapts to different decision contexts, and balances analytical sophistication with interpretability. This positions our framework as a practical blueprint for scalable, trust-worthy firm analytics in real-world settings.