Source & Coverage Proof

Data Sources

- **Primary**: SEC EDGAR CompanyFacts dataset (publicly available).
    - Downloaded from the SEC's official Company Facts API: https://www.sec.gov/search-filings/edgar-application-programming-interfaces
    - Contains entity-level financial disclosures in JSON format, including:
        - Net Income / Profit & Loss
        - Revenue / Sales
        - Assets, Liabilities, and Equity
        - Shares Outstanding, Public Float, and other standard GAAP metrics
- **Secondary**: yfinance (v0.2.65) for cross-checking select tickers with historical market data (prices, market cap).

Scope of Coverage

- **Entities**: All companies with a CIK registered in the CompanyFacts dataset at the time of download.
- **Metrics**: Financial statement items (net income, revenue, assets, liabilities, equity, shares outstanding, public float).
- **Time Period:** Includes all reported fiscal years and quarterly filings available in the dataset, typically spanning 10+ years of history per entity.
- **Granularity**: Each record is stored at company-year level with CIK, ticker, entity name, metric, value, and fiscal year.

Acquisition Log

- **Download Method:** Programmatic extraction via Google Colab notebook using Python (zipfile, os, json, tqdm).
- **Volume:**
    - Total JSON files processed: N (printed count from os.walk — e.g., ~6,000 files).
    - Total metrics extracted: M records (see df.shape output in notebook).

Data Cleaning & Transformation

The extraction pipeline applied multiple data normalization and quality steps:

1. **Tag Normalization**
    - Converted all SEC tag names to lowercase to avoid case-sensitivity mismatches.
    - Mapped hundreds of unique SEC tags into a smaller set of **semantic categories** (net income, revenue, assets, liabilities, equity, etc.) using keyword matching.
2. **Unit Handling**
    - Iterated through each available reporting unit (e.g., USD, shares) and selected numeric entries with valid `val` fields.
    - Ignored entries with null or missing values.
3. **Latest Value Selection**

- For each metric and reporting year, selected the **latest available record** (using `end` date) to avoid duplicates or overlapping periods.

4. **Entity Metadata Preservation**
    - Attached **CIK, ticker, and entity name** to every extracted record to allow cross-referencing with market data (via yfinance).
5. **Structured Output**
    - Built a structured list of dictionaries → converted to a **pandas DataFrame** for further analysis.
    - Ensured each row represents a single (company, metric, year, value) tuple for consistency.
6. **Category Completeness Check**
    - Stored the generated tag-to-category mapping as `TAG_MAP.json` for reproducibility and debugging.
    - Printed counts of matched tags per category to verify coverage.