Extracting and Exploring Causal Factors from Financial Documents

Thesis Presentation

Claire Zhao Sun

Heidelberg University
Institute of Computer Science
Database Systems Research Group
pz237@stud.uni-Heidelberg.de

July 13, 2022

Outline

1. Introduction
2. Problem Setting → What?
3. Design and Implementation → How?
4. Use Cases → Where?
5. Conclusion

Motivation

Text Mining Applications in Finance

- Scalability, efficiency and information advantage
- Sentiment analysis, anomaly detection, knowledge graph, etc.

Investment Research

- Information extraction and knowledge discovery
- Complex, multi-faceted process, comprehensive skill-set
- Mostly manual; lack of deployable text mining tools

Create a **text mining** tool to aid **investment research**

What does Investment Research entail?

- To identify unique investment opportunities
  - Understanding individual companies' business models
  - Following sector trends and industry dynamics: knowledge accumulation
  - Financial reports are a primary source of information
- Some specific questions in Investment Research:
  - What ***factors*** affect a company's financial performance?
  - How do these ***factors*** change over time? Are there any recognizable ***patterns***?
  - Which are the most ***closely related*** and comparable peers of a given company?
  - Given certain ***macro-level events***, which companies are most likely to be affected?

What are Causal Factors?

Definition

- A set of specific events, a general economic phenomenon, or a category of business activities, etc., which impact the financial performance of a company or a group of companies

- The scope can be specific or general, in order to capture:
  - immediate, direct drivers of key performance indicators (KPIs)
  - general shifts in macro and sector trends that affect many companies over the long run

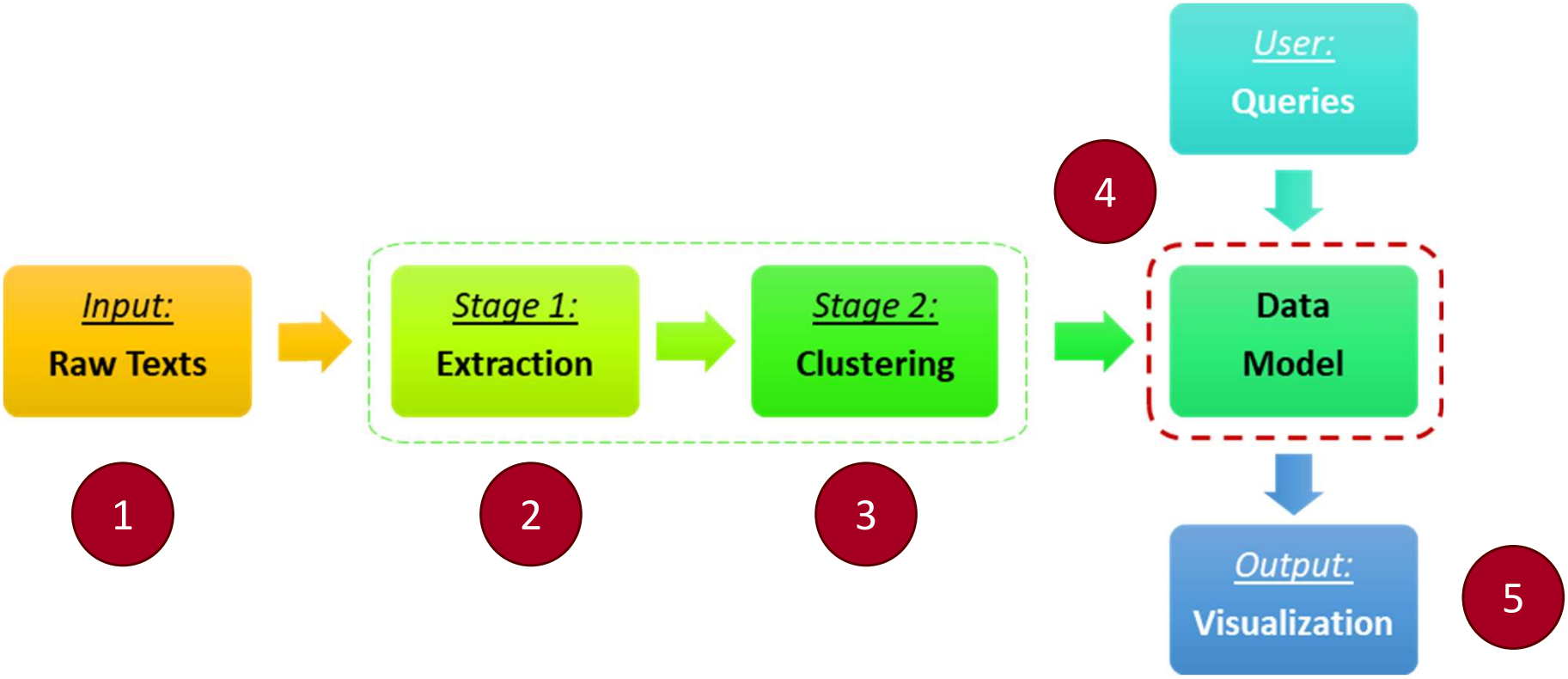
- Examples:
  - a pandemic such as COVID-19
  - wage increases
  - headcount changes
  - growth in paid subscriber base, etc.

Objectives

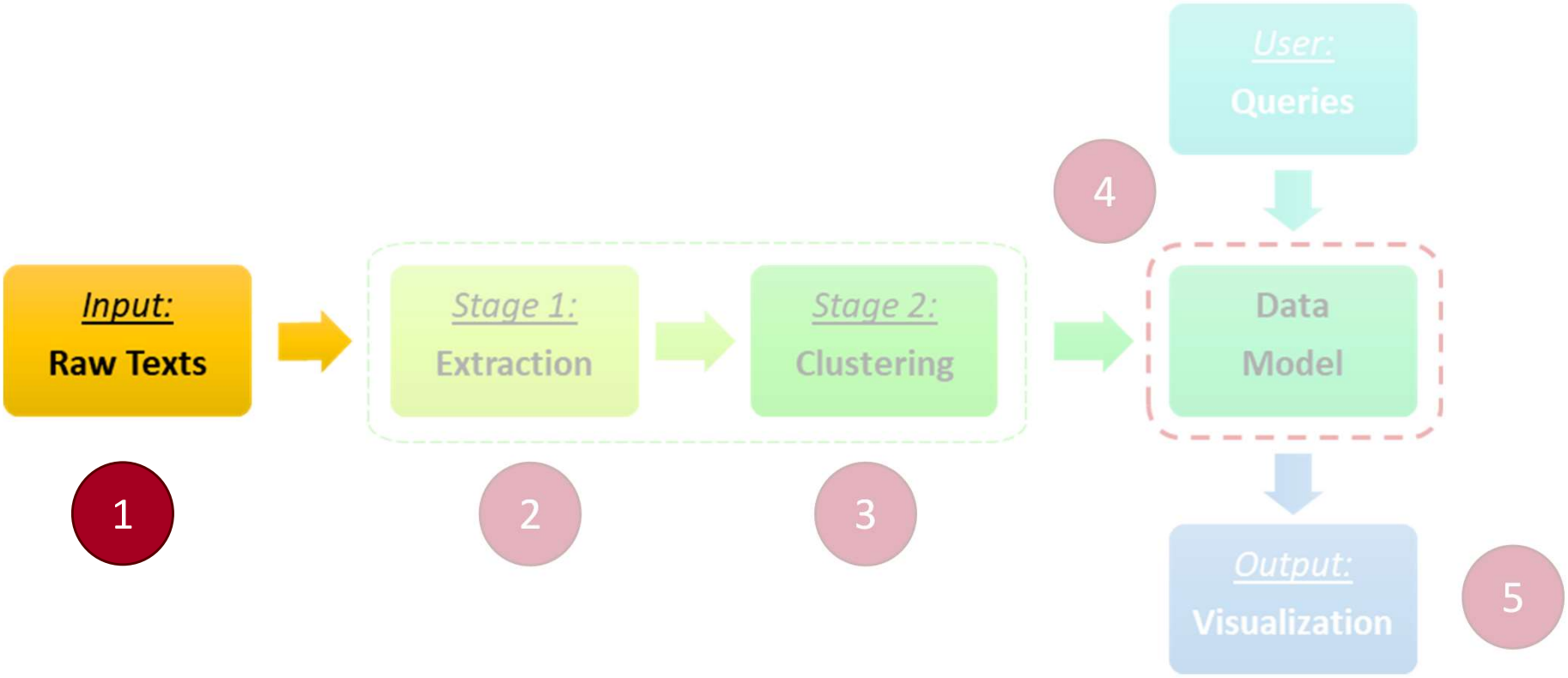
Create a **text mining** tool to aid **investment research**

- Extract **causal factors** from financial reports
- Design a **data model** that facilitates data exploration and knowledge discovery
- Implement an **end-to-end pipeline** that provides a **proof-of-concept** demo of solving real-life problems

Envisioned Pipeline



Envisioned Pipeline



1 Raw Data

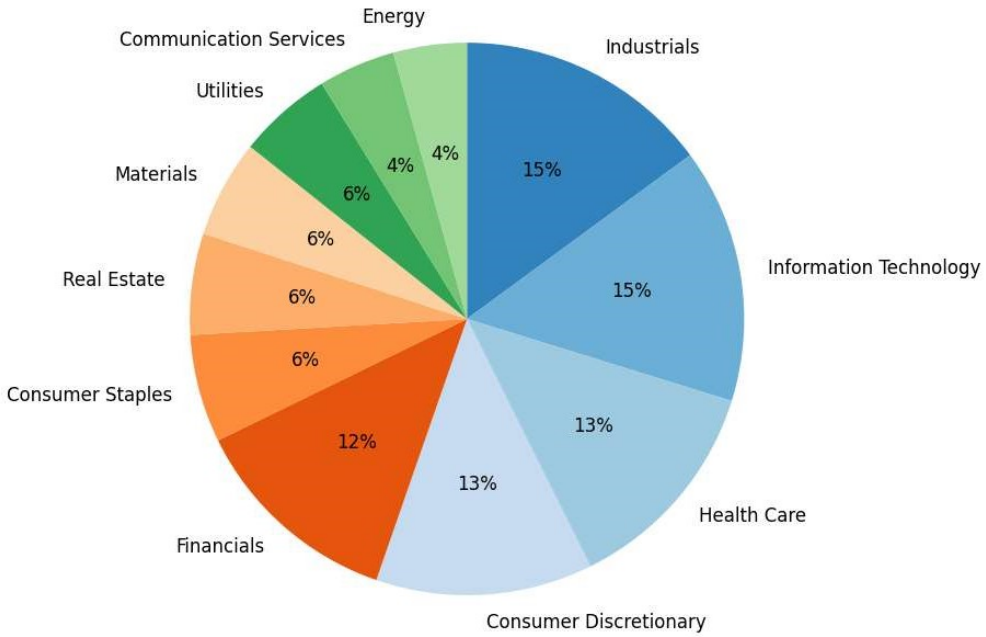
- Information Source
  - SEC filings
  - EDGAR database
  - Form-10K, Form-10Q
  - MD&A
- Data Scope
  - S&P 500 constituents
  - Last 20 years (2001 – 2021)
- Implementation
  - Download files
    - *sec-edgar-downloader*[1]
    - 40,628 files (HTML/txt)
    - Total size c. 85 GB
  - Extract MD&A
    - *BeautifulSoup*[2]
    - Use *regular expressions* to identify start and end positions based on section titles/headings
    - 32,807 MD&As extracted

Notes:

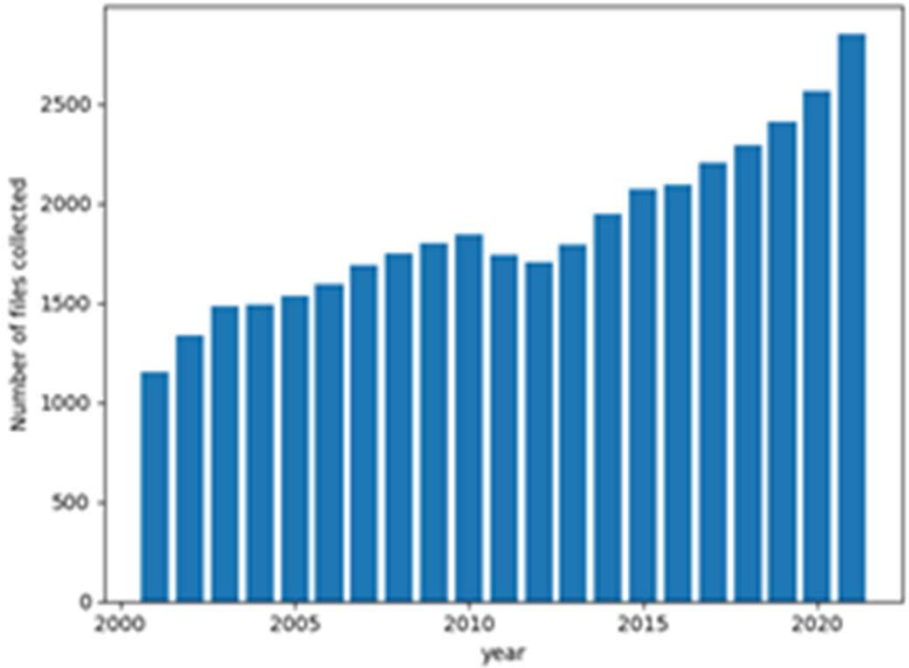
1. https://sec-edgar-downloader.readthedocs.io
2. https://beautiful-soup-4.readthedocs.io

1 Raw Data

Distribution of Companies by Sector

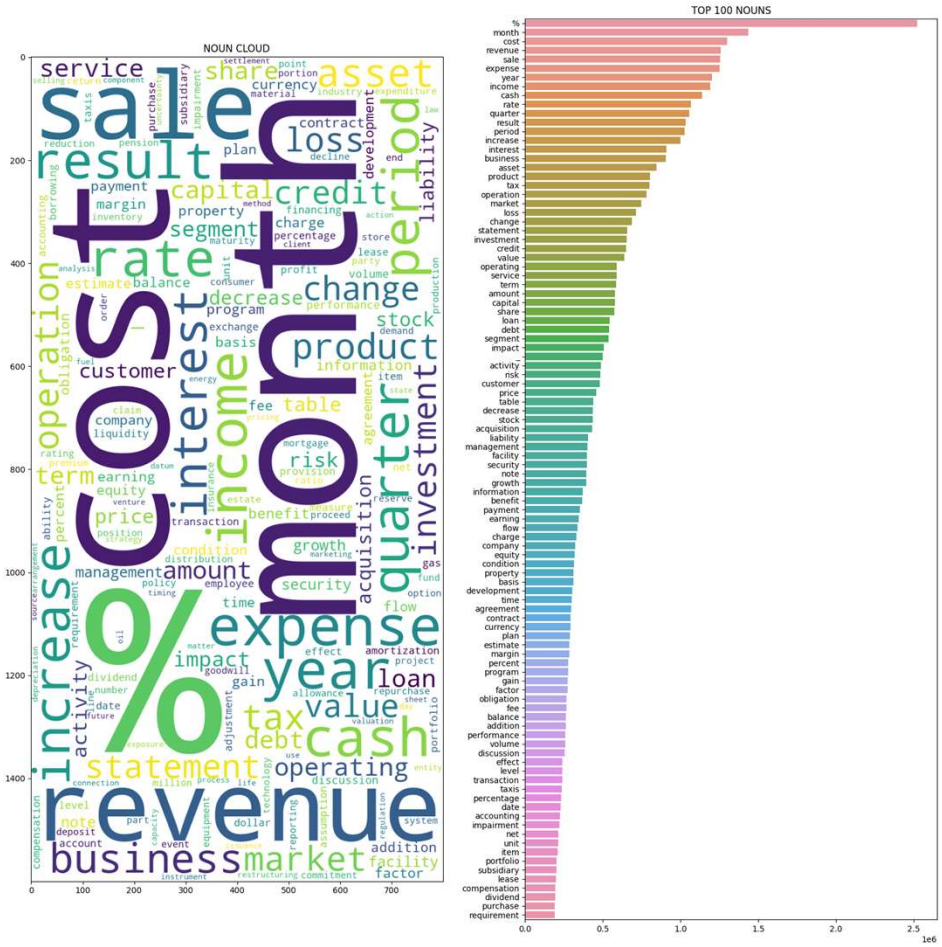


Distribution of Financial Reports by Year



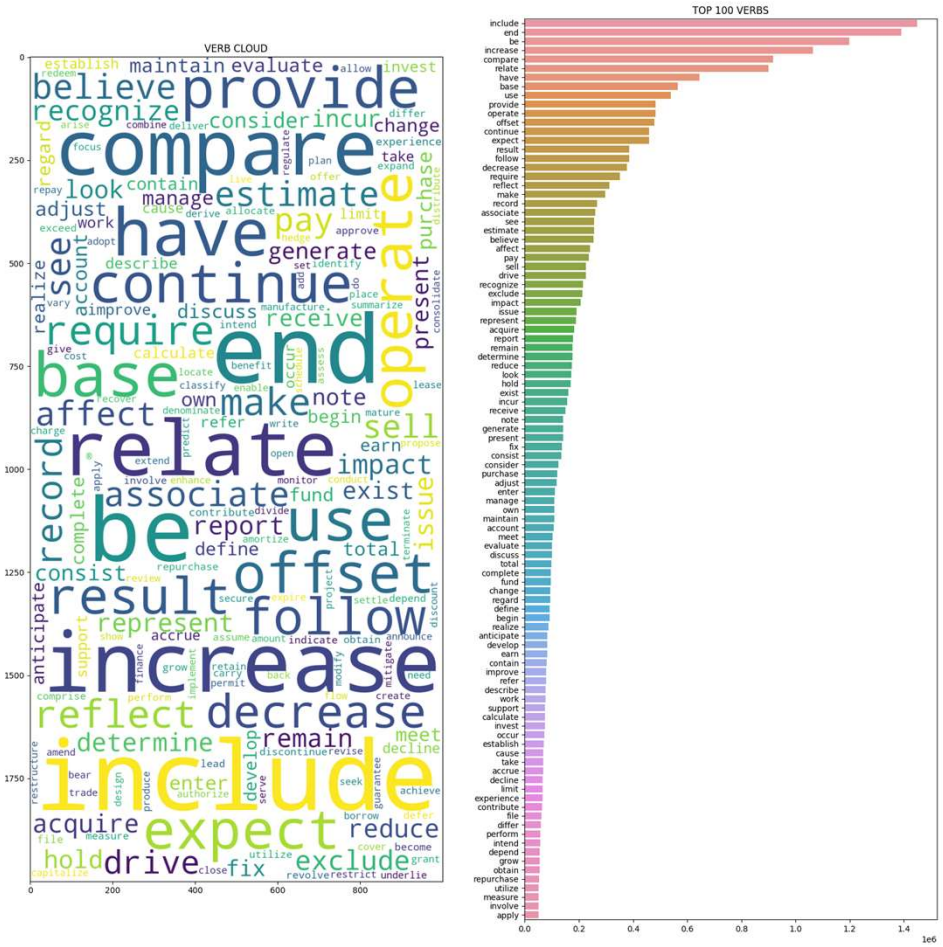
1 Raw Data

Top 100 Nouns in MD&A



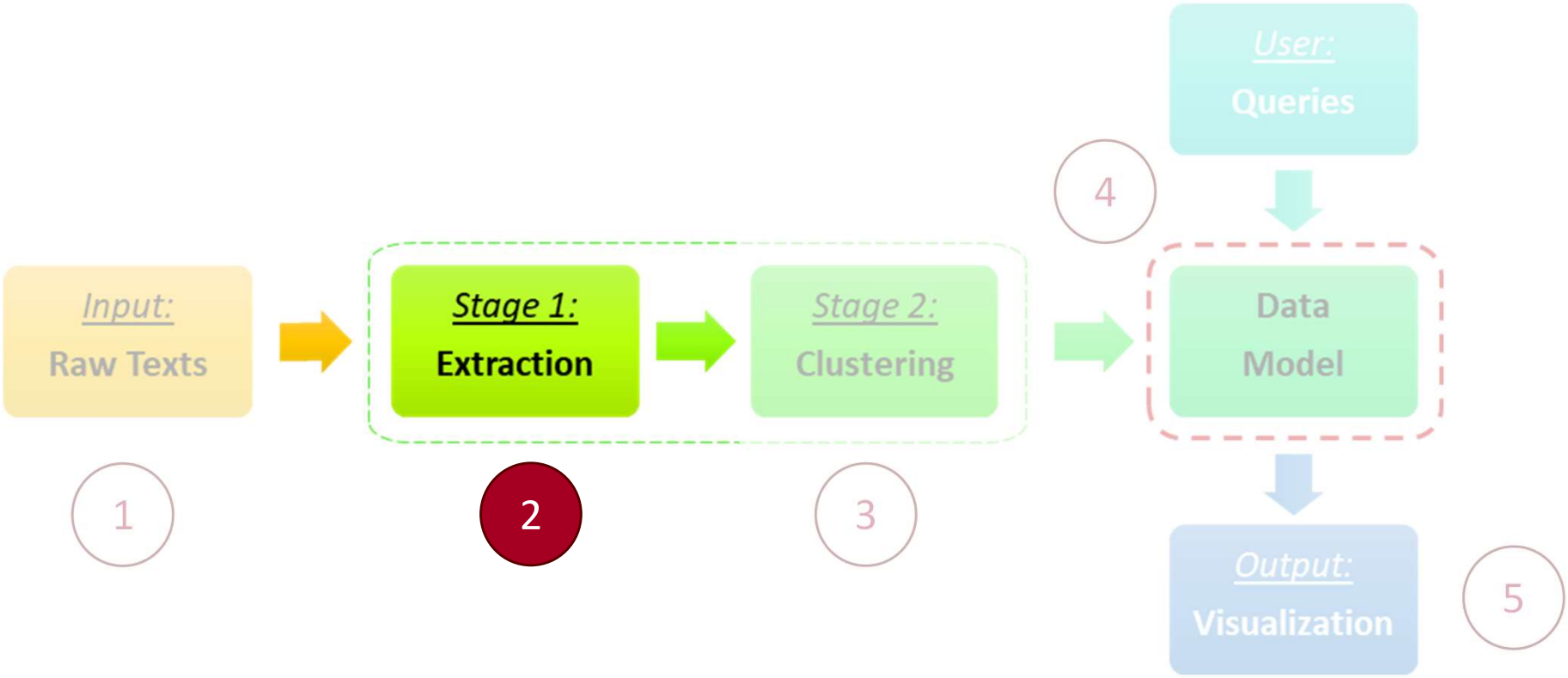
- Domain-specific, related to financial metrics

Top 100 Verbs in MD&A



- **Causal**: result, affect, impact, reflect, contribute, etc.

Envisioned Pipeline



2 Causality Extraction

Two Sub-tasks

- Classification: binary or multi-class (cause-effect, effect-cause, non-causal)
- Sequence tagging: BIO-scheme

Three Approaches

1. Rule-based linguistic patterns
   - Pre-defined causal words and/or lexico-syntactic patterns, e.g., <NP1, verb, NP2>
   - Semi-supervised learning
2. Traditional machine learning
   - SemEval datasets; classification only
   - Supervised learning
3. Neural network and deep learning
   - CNN, Bi-LSTM, BERT, etc.
   - Supervised learning

Paper	Year	Methodology	Data Source	Performance
Linguistic / Extraction				
Khoo [18]	1998	linguistic clues and pattern-matching	1,082 WSJ sentences	Recall 68% Precision 64%
Girju [21]	2003	lexico-syntactic pattern <NP1, verb, NP2>	LA TIMES section of the TREC 9 text collection	Recall 67% Precision 74%
Chan & Lam [22]	2005	- Semantic Expectation-based Knowledge Extraction framework; - Semantic template - WordNet-based component, used for finding concepts synonymous to the extracted case and effect phrases	News articles on topics such as Hong Kong stock market and global warming	*stock market* Recall 46% Precision 72% *global warming* Recall 56% Precision 64%
Kim et al. [15]	2007	linguistic connectives	50 aviation accident reports	Recall 67% Precision 72%
Radinsky et al. [16]	2012	- Textual causality patterns - event template for generalization	Over 14 million news articles	Recall 10% Precision 78%
Ittoo & Bouma [23]	2013	- weakly supervised system with Wikipedia as a knowledge-base - lexico-syntactic patterns based on a bootstrapping method	32,545 documents describing customer complaints and engineers' repair actions on medical equipment	Recall 82% Precision 77%
Traditional ML / Classification				
Rink et al. [25]	2010	SVM classifier based on lexical-syntactic features such as context words, hypernyms, POS, etc.	SemEval 2010 (1,331 causal sentences)	F1-score 78%
Sorgente et al. [24]	2013	Bayesian classifier based on lexical, semantic and dependency features	SemEval 2010 (1,331 causal sentences)	Recall 58% Precision 71%
Zhao et al. [26]	2015	Restricted Hidden Naive Bayes learning algorithm	SemEval 2010 (1,331 causal sentences)	F1-score 86%
DNN / Classification				
Kruengkrai et al. [27]	2017	Multi-column CNN	159,350 web sentences	Recall n.a. Precision 55%
Li & Mao [28]	2019	Knowledge-oriented CNN	SemEval 2010 (1,331 causal sentences)	Recall 90% Precision 93%
DNN / Extraction + Classification				
Dasgupta [29]	2018	Two layers of Bi-LSTM stacked, enhanced with an additional linguistic layer	Extended SemEval 2010 (8,000 sentences)	*SemEval F1-socre* Cause 84% Effect 79% Connectives 75%
Chen et al. [30]	2020	2-layers of Bi-LSTMs (sentence segmenter + relation classifier)	69,120 manually labelled sentences	Recall 93% Precision 92%
Li et al. [17]	2020	BiLSTM with multi-head self-attention	Extended SemEval 2010 (4,450 sentences)	Recall 83% Precision 86%
Kao et al. [32]	2020	BERT-CRF system and a Viterbi decoder for span optimization	FinCausal 2020 (1,579 causal sentences)	F1-score 95%
Becquin [33]	2020	BERT-SQUAD augmented system with heuristics for span	FinCausal 2020 (1,579 causal sentences)	F1-score 95%

2 Causality Extraction

Considerations

1. Rule-based Approach
   - ✓ Simple, transparent and adjustable
   - × Manually constructed patterns
   - × Difficult to capture complexity of causal expressions

2. Traditional ML Approach
   - × Classification only with labelled datasets
   - × Rely on feature engineering and extraction tools
   - × Performance not necessarily better than 1

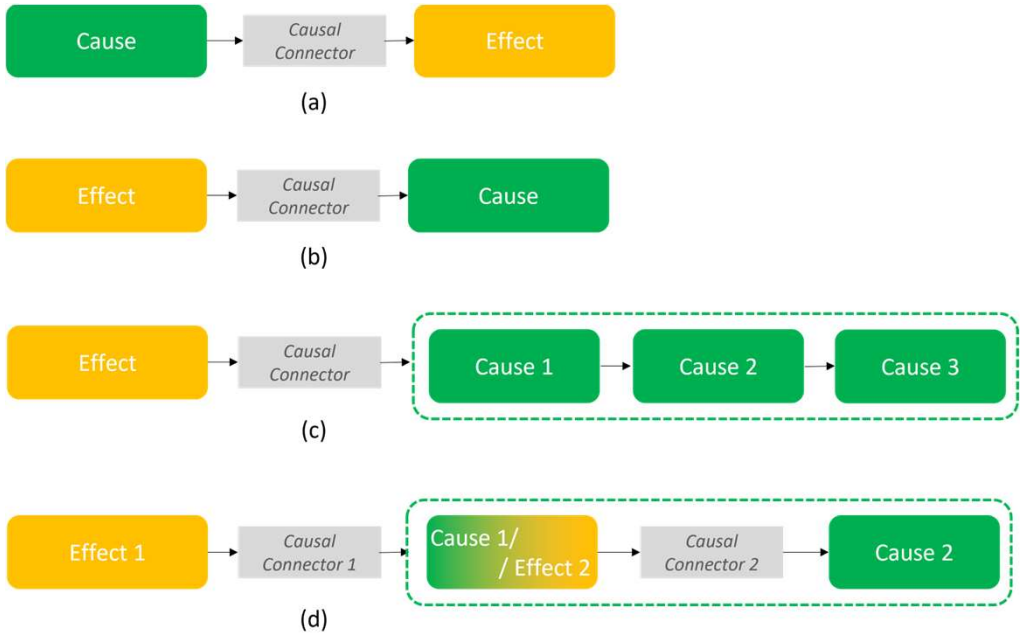
3. Deep Neural Network Approach
   - ✓ Best Performance
   - × Large labelled dataset required for training
   - × Limitations of currently available datasets
   - × Complexity and computationally intensive

Implementation

- Task at hand:
  - Identify causal sentences
  - Identify cause and effect chunks
  - Extract noun phrases
- Understanding MD&A language
  - Domain-specific, business formal
  - Credit attribution, logical
  - Sentence structure patterns
- Rule-based approach
  - Practicality considerations for deployment
  - Simplicity triumphs complexity

2 Causality Extraction

Causal Sentence Patterns



- **Causal Sentence:** a sentence containing three components: a cause (C), an effect (E) and an explicit causal connector (CC)
- **Causal Connector:** a linguistic signal of causality, can be verb (e.g., E *driven by* C), prepositional phrase (e.g., E *due to* C), conjunction (e.g., E *because* C), etc.
- **Cause Chunk:** a text fragment describing an event or phenomenon that *claims* to cause another event or phenomenon, which is described by the **Effect Chunk**
- Any causal sentence can be represented as one of **two patterns** of cause-effect chunks linked by causal connectors (see illustration (a) and (b))

Examples

- **Multiple causes:** "[Growth for our direct response advertising products] was primarily *driven by* **[increased advertiser spending]** as well as **[improvements to ad formats and delivery]**."
- **Multiple connectors:** "[There was an increase in operating expenses] primarily ***driven by*** [an increase in compensation expenses] largely ***due to*** [increases in headcount]."

2 Causality Extraction

Implementation

- Causal connectors
  - **Verbs:** *result, affect, impact, cause, contribute, attribute, drive, relate, associate, reflect*
  - **Non-verbs:** *due to, because, attributable to, as a result of, in connection with*
- Effect chunks
  - **Topics:** revenue, sale, cost, expense, margin, profit, earning, income, loss
- Cause chunks
  - **Noun phrases:** a noun plus the words describing the noun
- SpaCy[1]
  - Sentence segmentation, rule-based matching, syntactic dependency parser, etc.

Results

- **1,003,152** causal sentences
- **5,069,900** noun phrases
- **455,538** *unique* noun phrases

Notes:

1. https://spacy.io/

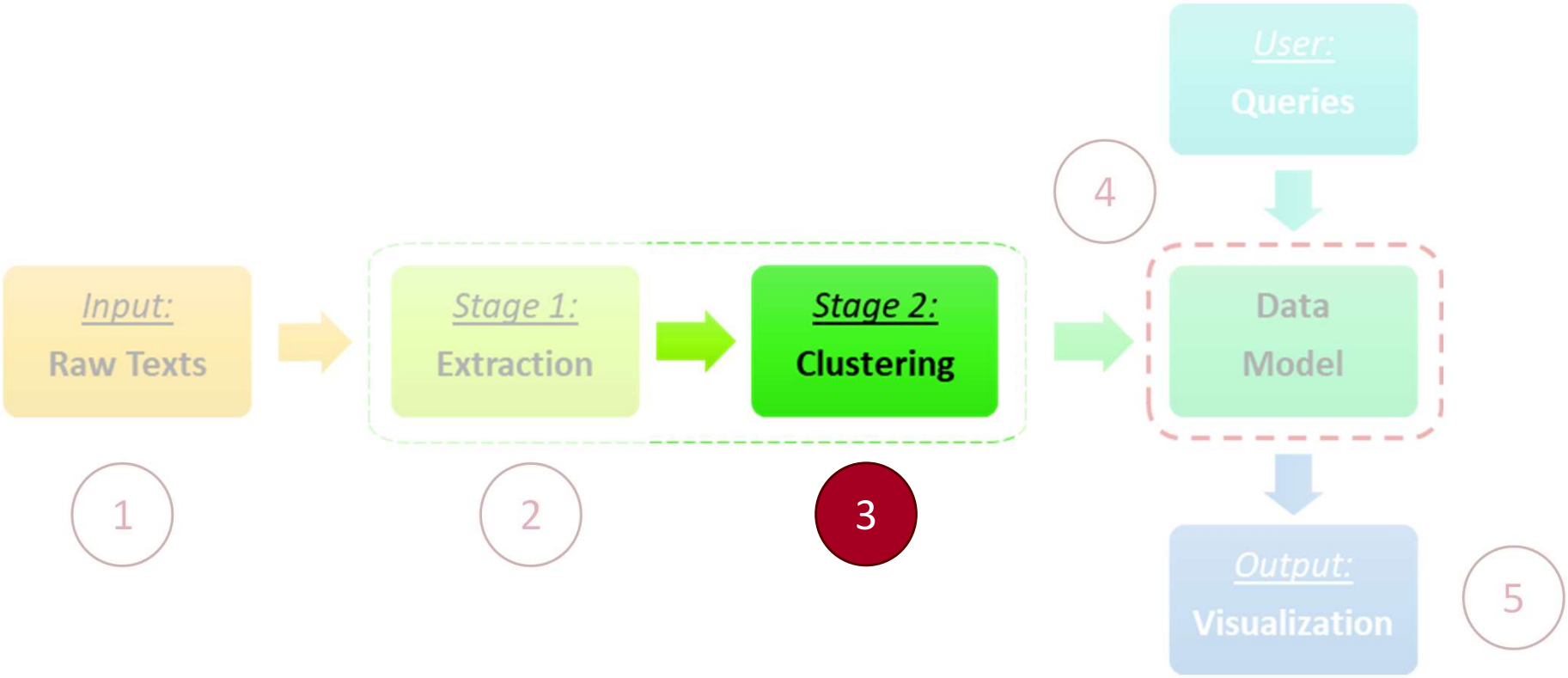
2 Causality Extraction

Evaluation

- Manually annotated sentences from the three raw MD&As
  - Binary classification: causal vs non-causal
- Module performance: precision 90%, recall 89% and F1-score 90%
  - State-of-the-art Bi-LSTMs: precision 92%, recall 93% [30]
- Limitations and improvement
  - Not statistically robust (sample size too small); not meaningful to benchmark against a publicly available general-purpose dataset
  - More time resources and/or external help for annotation

Doc	#Casual Sentences Manually Identified	#Causal Sentences Extracted by Model	#TP	#FP	#FN	P%	R%	F1%
1	98	94	88	6	10	94%	90%	92%
2	80	80	74	6	6	93%	93%	93%
3	74	76	63	13	11	83%	85%	84%
All	**252**	**250**	**225**	**25**	**27**	**90%**	**89%**	**90%**

Envisioned Pipeline



3 Clustering

Purpose

- Noun phrases → Semantic groups

Embeddings

- Word-level: Word2Vec, GloVe
- Character-level: ELMo, Flair
- Contextualized: BERT and variants

Algorithms

- K-means
  - Preset number of clusters; no visual output
- Hierarchical Clustering
  - Time $O(N^2 \log N)$ vs. $O(NM)$; Space $O(N^2)$ vs. $O(N+M)$
- Self-Organizing Map
  - 2D visualization

Implementation

- Phrase Embeddings:
  - Fin200: Finance-specific word embeddings trained with the continuous bag-of-words (CBOW) model over 40,000 10-K filings [95]
  - Simple average, normalized to a range of [-1,1]
- Self-Organizing Map
  - MiniSOM 2.3.0[1]
  - Topology: 58x58 grid
  - Neighbourhood function: Gaussian
  - Sigma and learning rate determined through experimentation
- Hierarchical Clustering *(Optional)*
  - Scikit-learn 1.1[2]: Agglomerative Clustering model
  - Number of clusters: Elbow method with Silhouette Scores as a measure of cluster quality

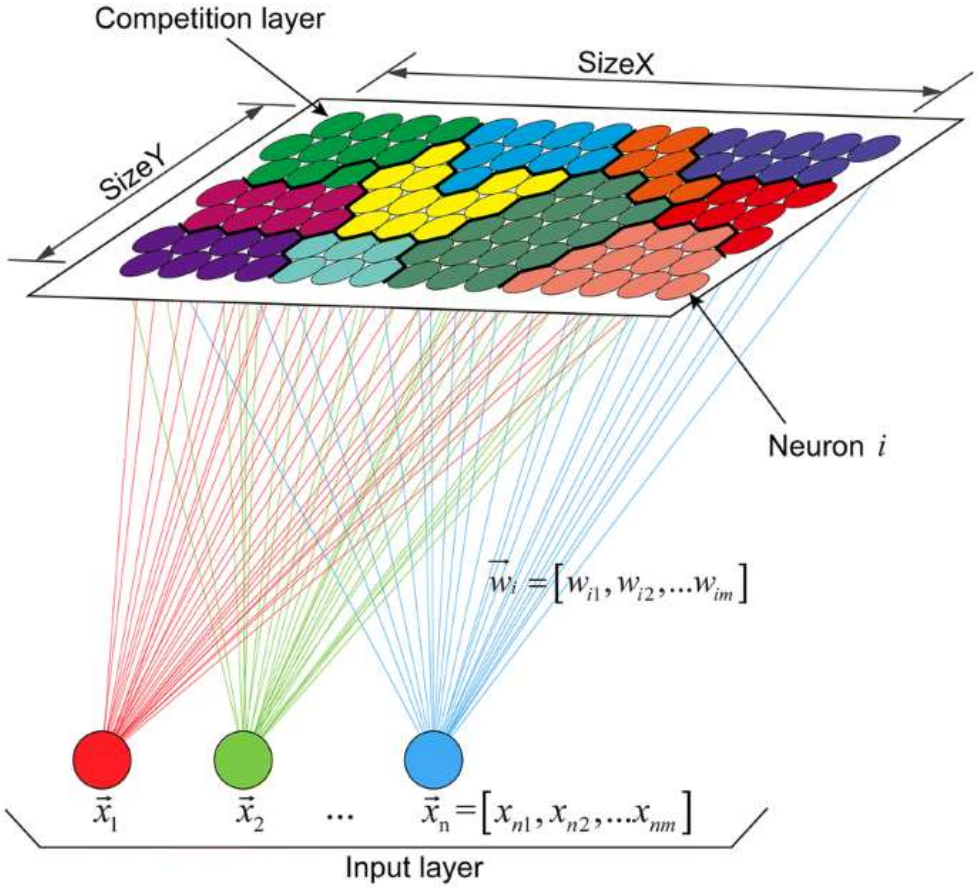
Notes:

1. https://pypi.org/project/MiniSom/
2. https://scikit-learn.org/

3 Clustering

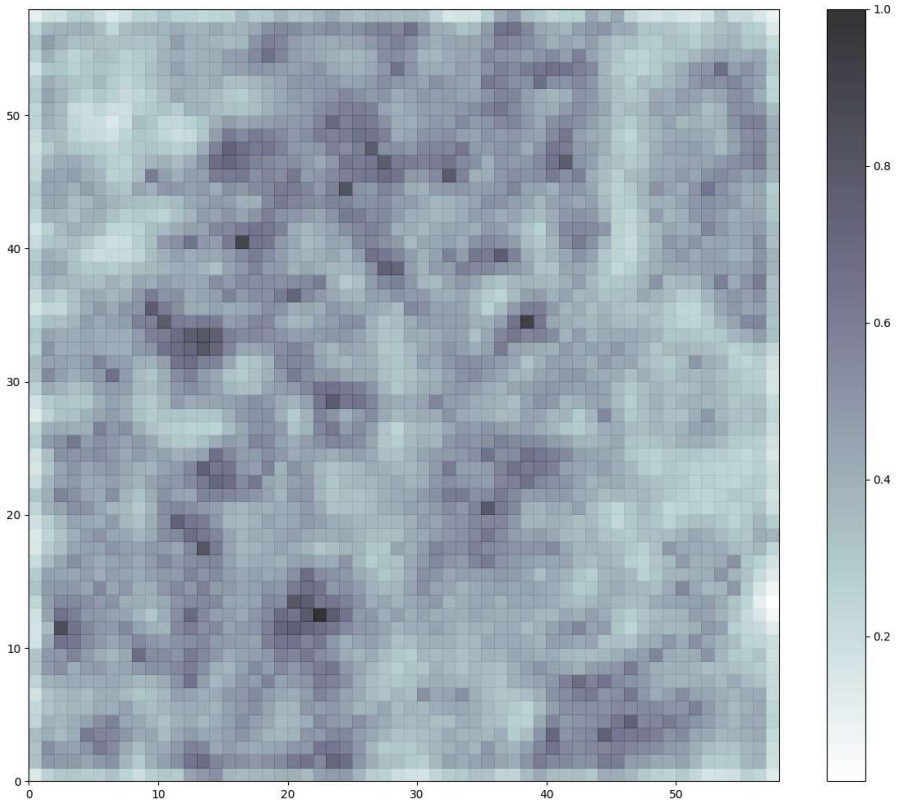
Self-Organizing Map (SOM)

- Feed-forward neural network trained with a competitive learning algorithm
- Takes *n*-dimensional input data and outputs a *2D* map, where regions can be interpreted as clusters and their geometric closeness indicating similarity
  - Often used as a data visualization technique
- Steps of the Algorithm:
  1. ***Initialize:*** Select grid size, topology, neighborhood function and learning rate; neurons initialized with random weights (same dimension as input)
  2. ***Sample:*** Select an input data point at random and feed it into the SOM
  3. ***Match:*** Find the Best Matching Unit (BMU), i.e., the neuron whose weight most similar to the input data (usually Euclidean distance)
  4. ***Update:*** Adjust the weight of the BMU and its neighbors to be closer to the input data according to the preset neighborhood function and learning rate
  5. ***Repeat*** steps 2 - 4 until the neuron weights stabilize



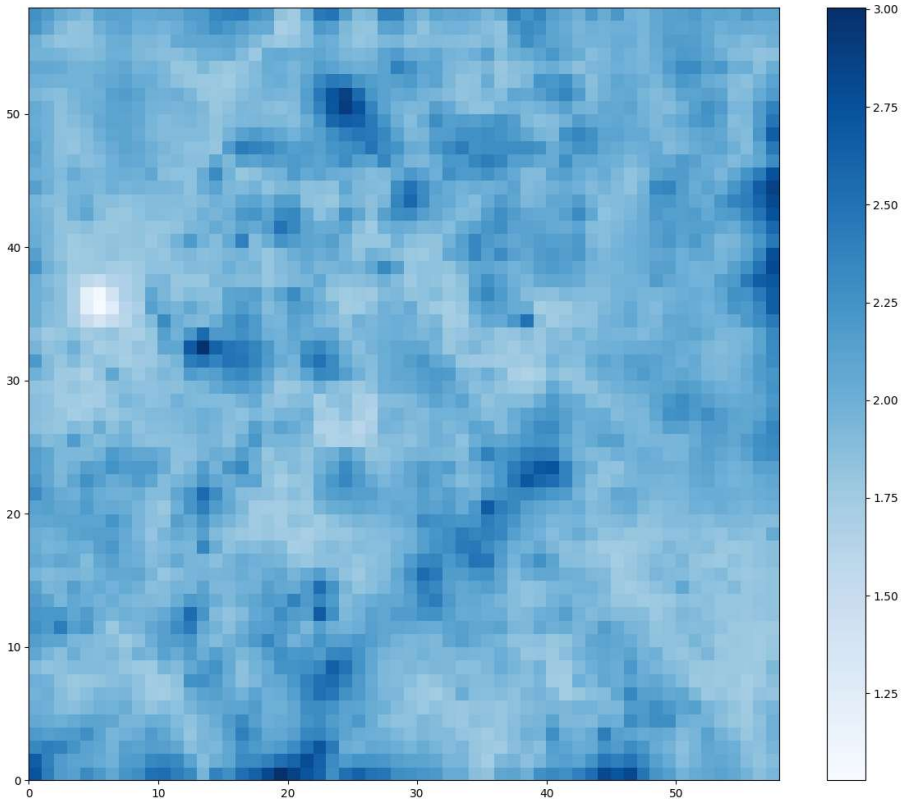
3 Clustering

SOM Distance Map



Visualization of how similar each neuron is to its neighbors. The color intensity corresponds to the normalized sum of the Euclidean distances of each neuron's weight to its neighbors.

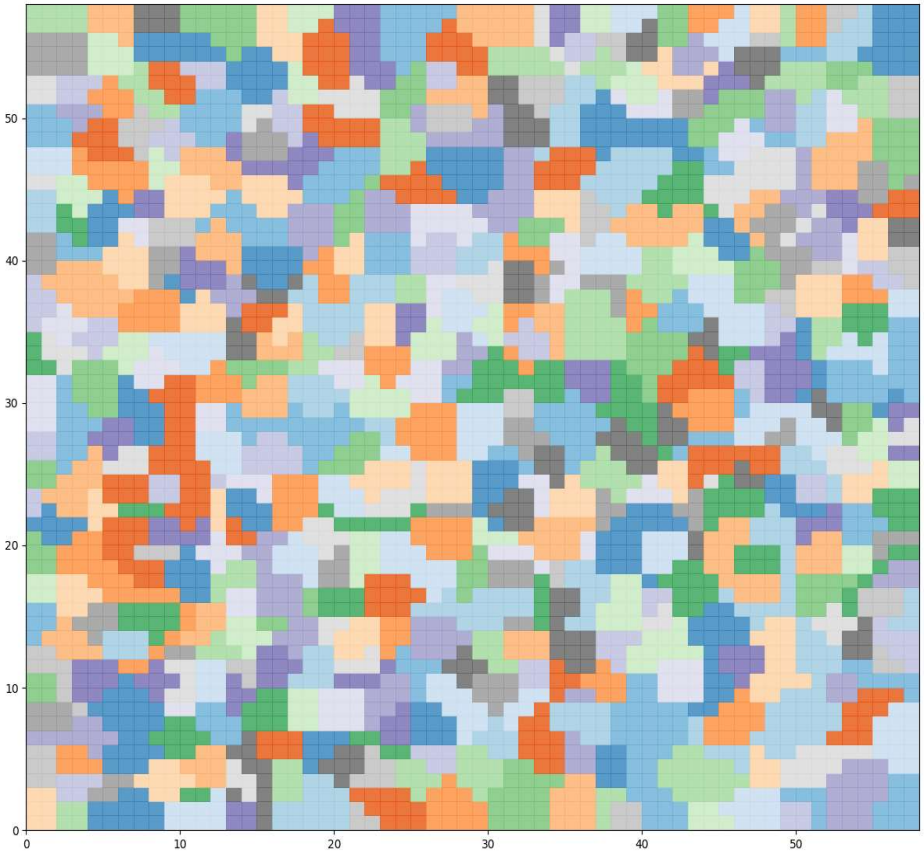
SOM Activation Map



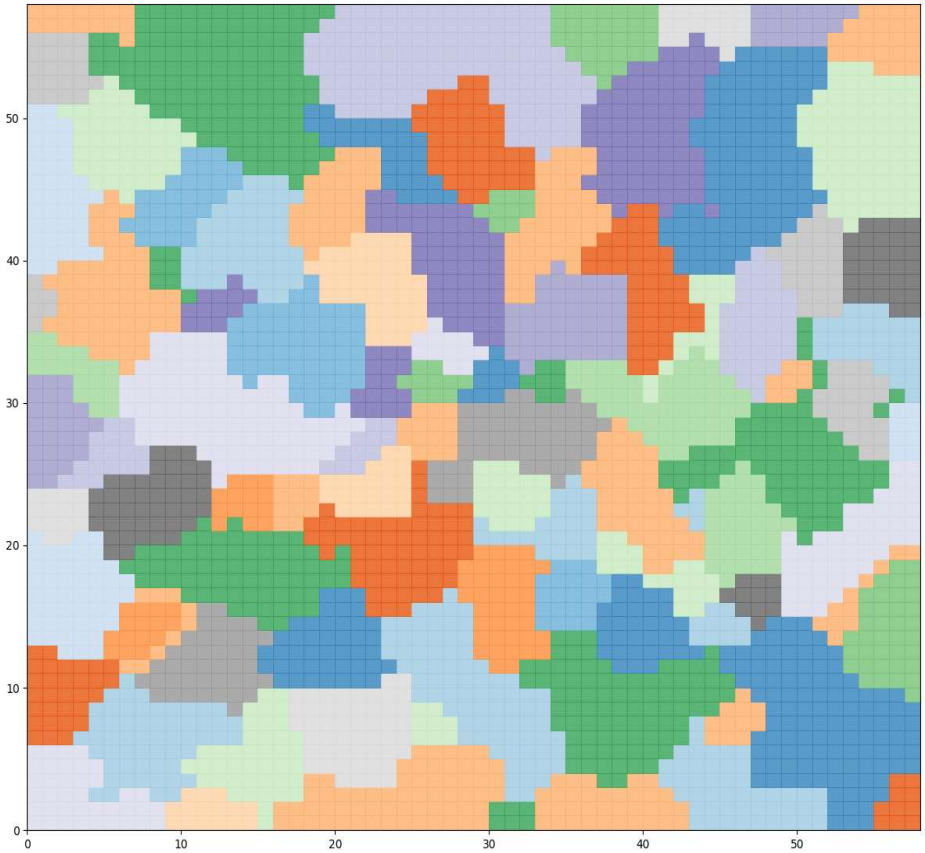
Matrix representation of the neuron network where the color intensity corresponds to the number of times that underlying neuron has been updated.

3 Clustering

Map of 500 H-Clusters



Map of 50 H-Clusters



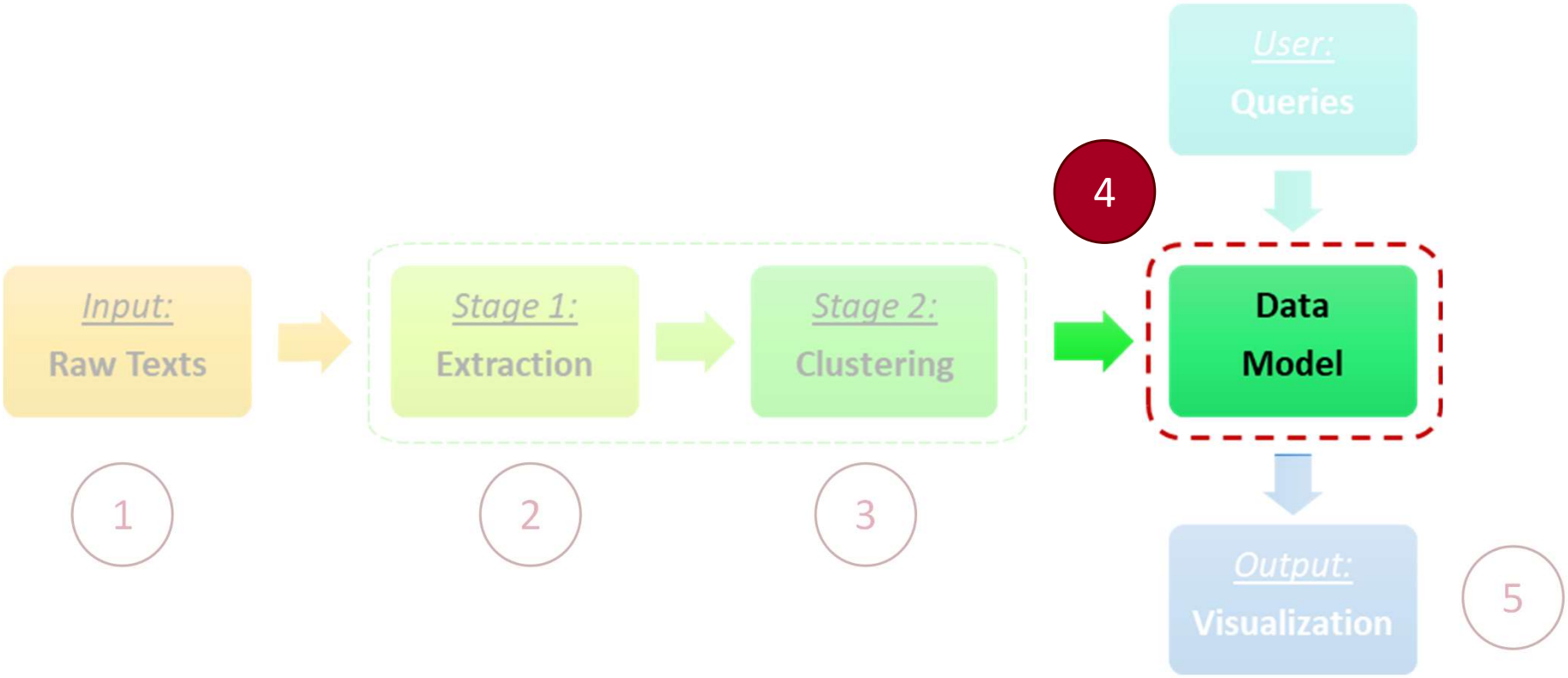
3 Clustering

Evaluation

- No objective ground truth
  - Quantitative metrics insufficient to fully describe true quality of clusters (inherently subjective)
- Qualitative judgement
  - Random sampling: 10 H-clusters, 5 SOM-clusters, 10 sample NPs
  - Each SOM-cluster consists of semantically similar NPs
  - H-clusters merge similar SOM-clusters together
  - The underlying neurons in close proximity on the 2D SOM map
  - Abstraction by scale and scope
- Potential improvement
  - Recruit subject area experts' help

Cluster ID	Sample Noun Phrases
H-Cluster 471	
SOM #924	'rapidly evolving macroeconomic environment', 'improved market environment', 'stable market environment', 'favorable lending environment', 'slow sales environment', 'recessionary US environment', etc.
SOM #925	'positive economic outlook', 'continue economic strength', 'challenge consumer economic and geopolitical environment', 'volatile economic environment', etc.
SOM #926	'economic downturn', 'continue global economic weakness', 'global macroeconomic challenge', 'weak economic datum', 'slow economic recovery', 'macroeconomic decline', etc.
SOM #982	'current covid19 environment', 'broad base economic recovery', 'improved macroeconomic outlook', 'certain macro economic and business factor', 'global macro hedge fund', 'macro related product', etc.
SOM #984	'overall economic climate', 'economic stimulus payment', 'difficult local economic', 'numerous economic and political issue', 'domestic and international economic and political consideration', etc.
H-Cluster 376	
SOM #1624	'low oil gas and NGL', 'natural gas throughput volume', 'net wellhead oil and natural gas price', 'natural gas and crude oil financial price swap contract', 'high natural gas and NGL price', etc.
SOM #1683	'physical natural gas cost', 'natural aging', 'its natural gas liquid', 'natural gas sale net', 'broker natural gas revenue', 'physical and forward natural gas', 'its natural gas distribution', etc.
SOM #1741	'natural gas and power contract', 'its natural gas system', 'natural gas cost recovery mechanism', 'under recover natural gas revenue', 'gas supply contract', 'natural gas backup generator', 'power and natural gas market', etc.

Envisioned Pipeline



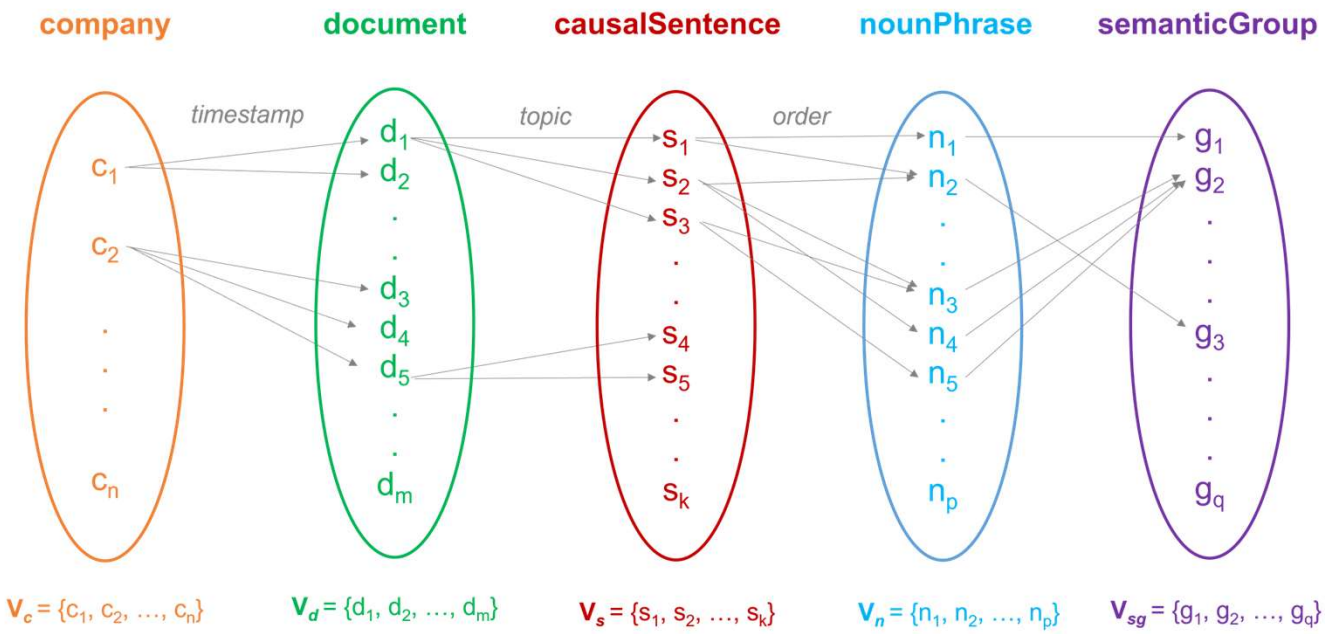
4 Data Model

Purpose

- Structured way of storing causal information extracted from financial reports
- Means of connecting companies via semantic clustering of causal factors

Key Characteristics

- Heterogeneous graph with 5 node types:
  - company, document, causalSentence, nounPhrase and semanticGroup
- Directed edges
  - Indicating order of hierarchy
  - Storing temporal information
- Dynamic in nature
  - Incrementally updated

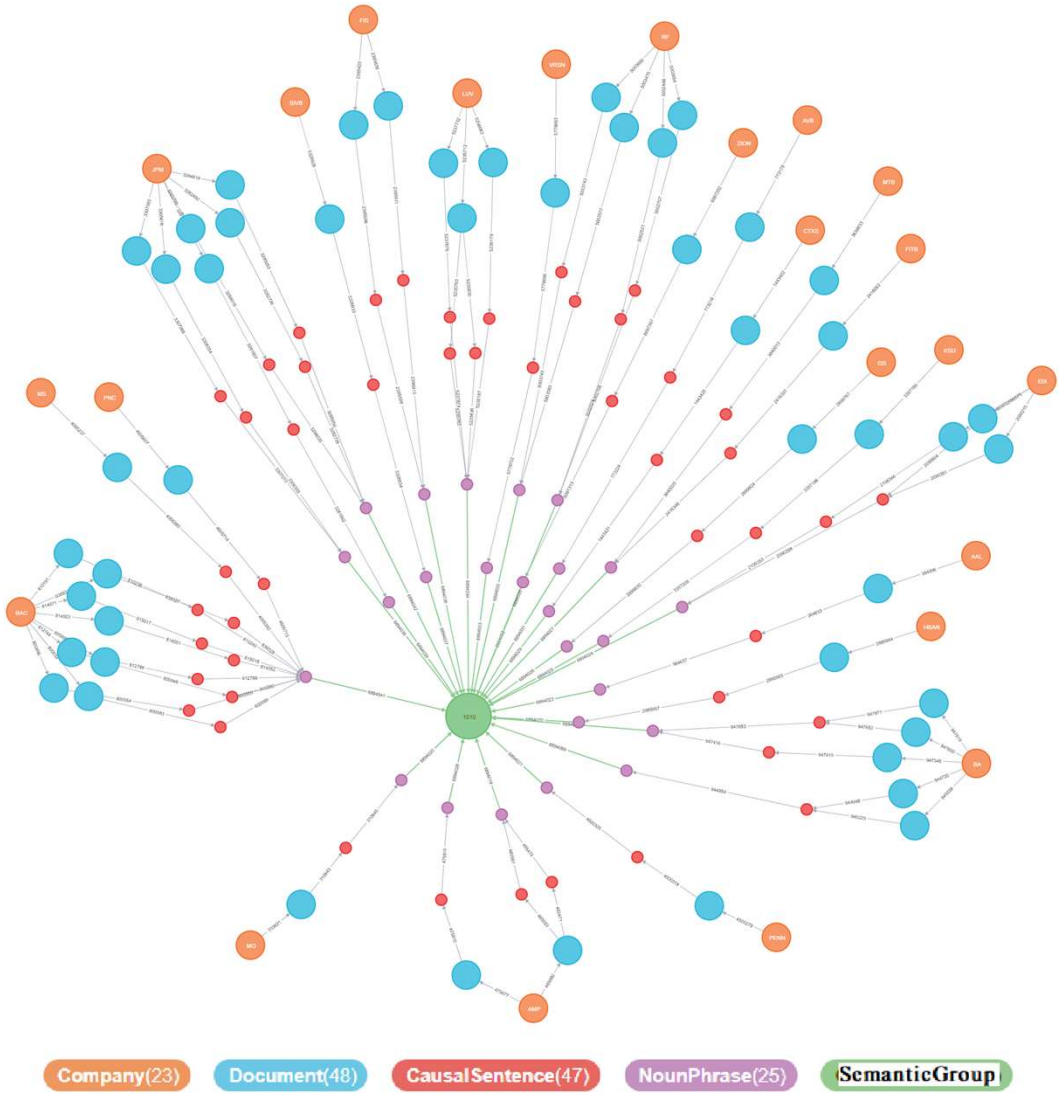


4 Data Model

Implementation

- Neo4j Community Edition 4.0.4[1]
  - 500 company nodes
  - 32,807 document nodes
  - 1,003,152 causalSentence nodes
  - 455,538 nounPhrase nodes
  - 3,359 semanticGroup nodes *(before the optional second-stage hierarchical clustering)*

- Visualization of data model
  - A sample cluster around a semanticGroup node (center, green)
  - Nodes in the outer layers: nounPhrase (inner-most, purple), causalSentence (red), document (blue) and companies (outer-most, orange)

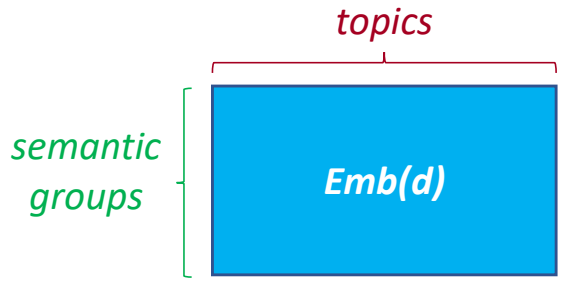


Notes:
1. https://neo4j.com/

4 Data Model

Document Nodes

- $Emb(d) = [u_{i,j}] \in \mathbb{R}^{m \times n}$
  - d is a node of type *document*
  - $u_{i,j}$ = count of semanticGroup *i* under topic *j*
  - *m* = Total number of *semanticGroups*
  - *n* = Total number of *topics*



- $Sim(d_A, d_B) = [\boldsymbol{s_j}] \in \mathbb{R}^{1 \times n}$

$$\boldsymbol{s_j} = \frac{\boldsymbol{a_j} \cdot \boldsymbol{b_j}}{\|\boldsymbol{a_j}\| \cdot \|\boldsymbol{b_j}\|}$$

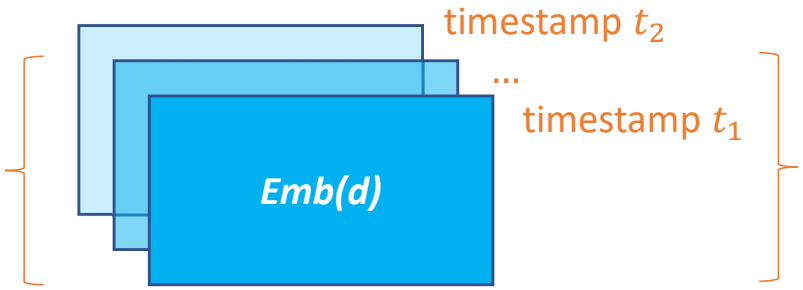
$$\boldsymbol{a_j} = Emb(d_A)[:, j]$$

$$\boldsymbol{b_j} = Emb(d_B)[:, j]$$

  - The resulting representation is a row vector with each entry representing the similarity measure for each topic

Company Nodes

- $Emb\left([c]_{t_1}^{t_2}\right) = \sum_{t_1}^{t_2} Emb(\mathrm{d}_t) \mid d_t \in \boldsymbol{\mathcal{N}}(c)$
  - c is a node of type *company, the node embedding is specified with the time period from t_1 to t_2*
  - d_t = node of type *document,* with timestamp *t*
  - $\boldsymbol{\mathcal{N}}(c)$ = neighbourhood of node *c*



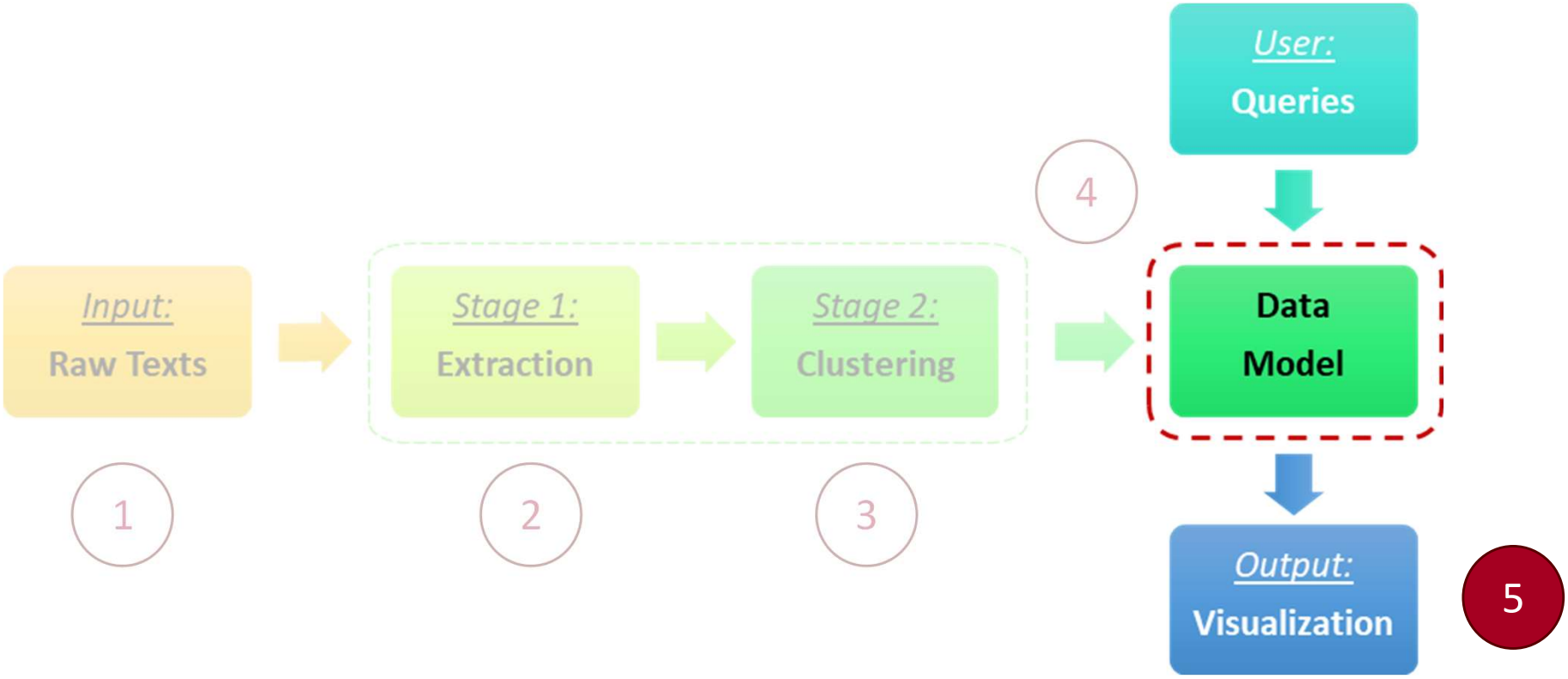
- $Sim\left([c_A]_{t_1}^{t_2}, [c_B]_{\tau_1}^{\tau_2}\right) = [\boldsymbol{s_j}] \in \mathbb{R}^{1 \times n}$

$$\boldsymbol{s_j} = \frac{\boldsymbol{a_j^t} \cdot \boldsymbol{b_j^\tau}}{\|\boldsymbol{a_j^t}\| \cdot \|\boldsymbol{b_j^\tau}\|}$$

$$\boldsymbol{a_j^t} = Emb\left([c_A]_{t_1}^{t_2}\right)[:, j]$$

$$\boldsymbol{b_j^\tau} = Emb\left([c_B]_{\tau_1}^{\tau_2}\right)[:, j]$$

Envisioned Pipeline



What does Investment Research entail?

- To identify unique investment opportunities
  - Understanding individual companies' business models
  - Following sector trends and industry dynamics: knowledge accumulation
  - Financial reports are a primary source of information

- Some specific questions in Investment Research:
  - What ***factors*** affect a company's financial performance?
  - How do these ***factors*** change over time? Are there any recognizable ***patterns***?
  - Which are the most ***closely related*** and comparable peers of a given company?
  - Given certain ***macro-level events***, which companies are most likely to be affected?

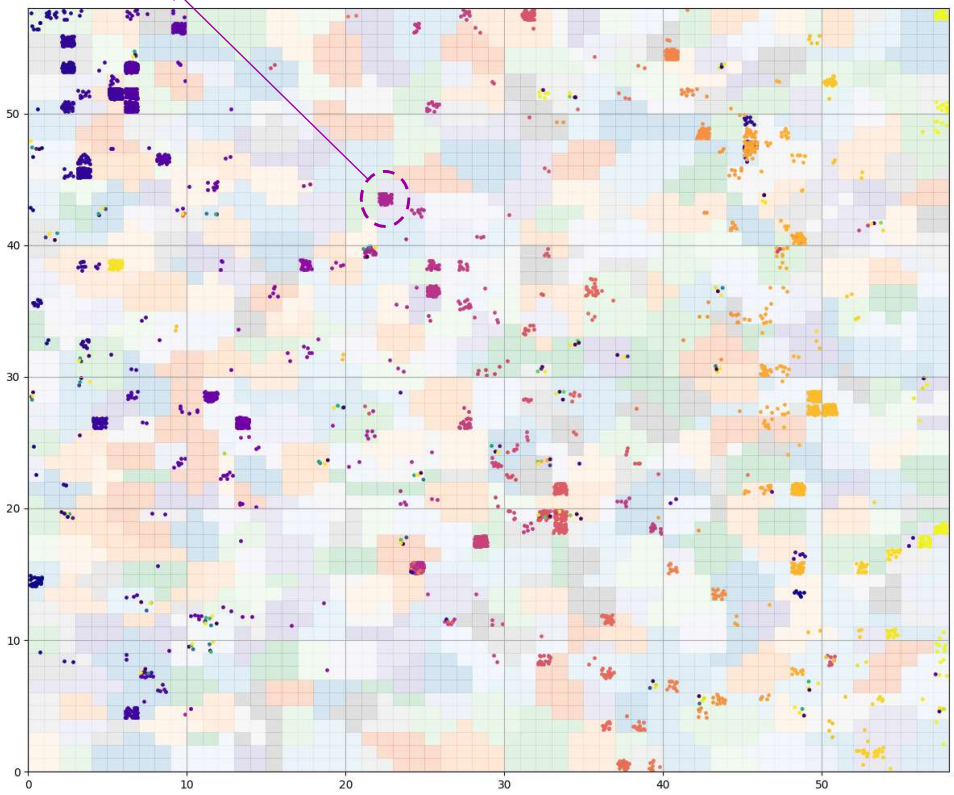
5 Use Cases

- Comparison of Companies' Visual Signatures

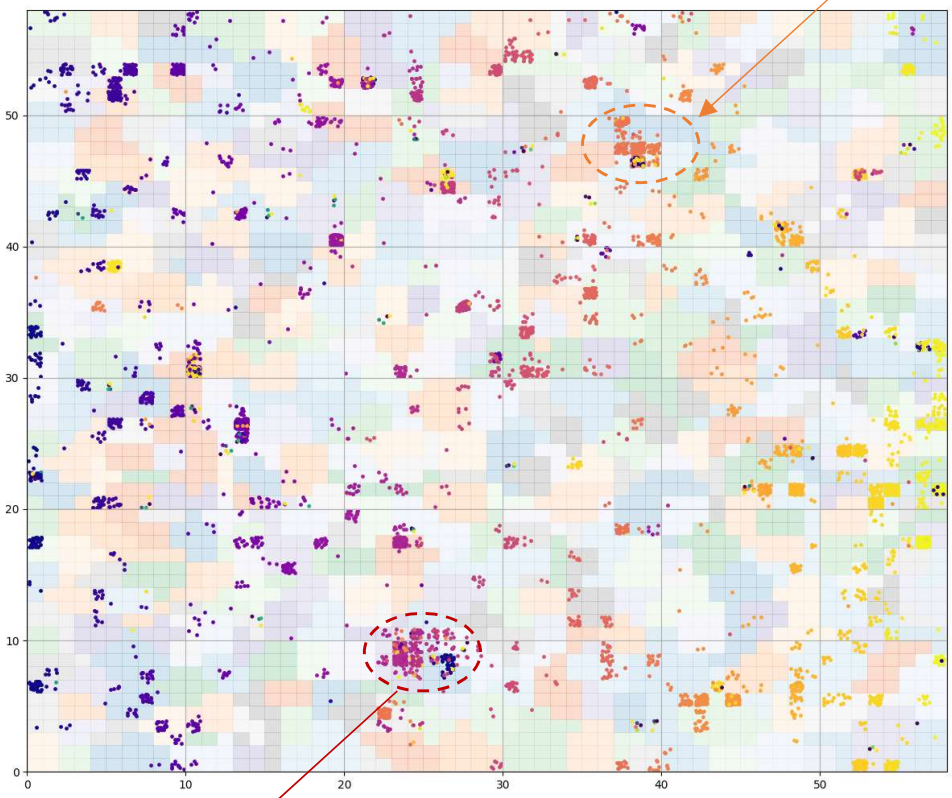
Brand identity, brand loyalty, etc.

Multi-year Causal Factor Maps (2002-2021)

Production cost, efficiency improvement, optimization, etc.



Netflix



Engineering, manufacturing service, installation, etc.

Boeing

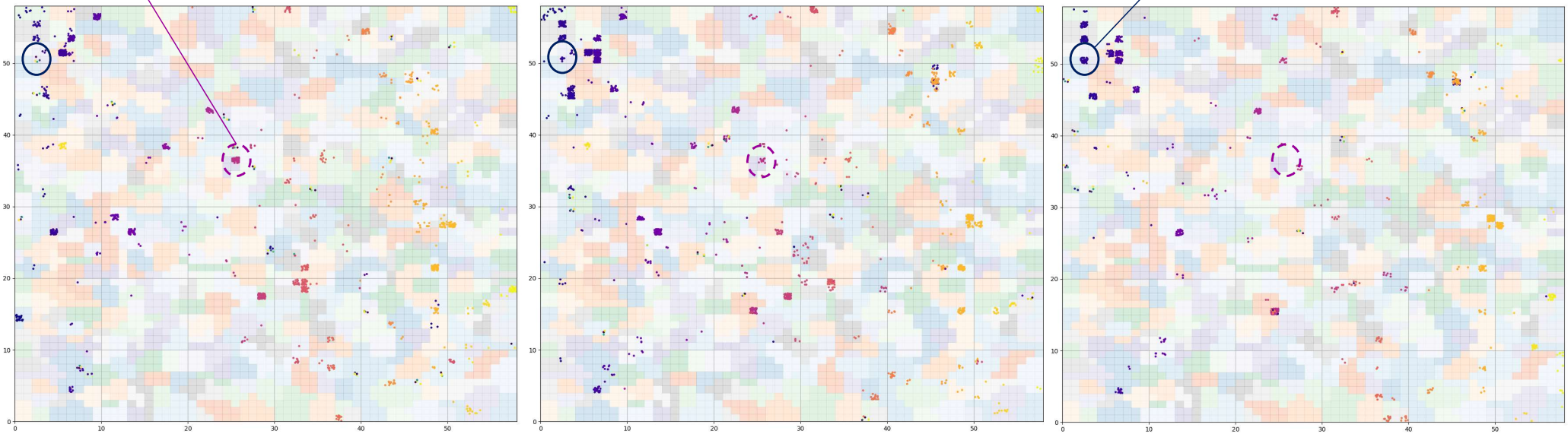
5 Use Cases

- Displaying the Evolution of Causal Factors through Time

Brand identity, brand loyalty, etc.

Netfix's Multi-year Causal Factor Maps

Customer experience, etc.



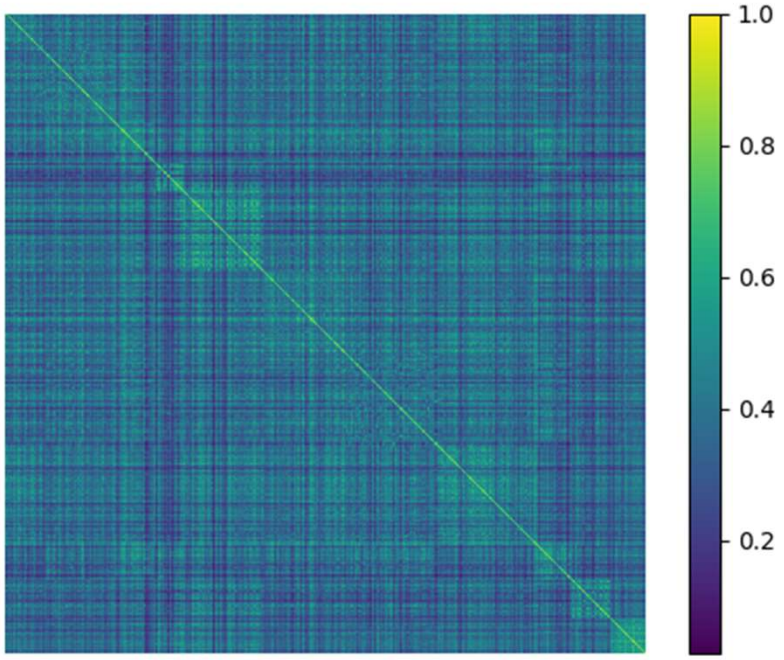
2006-2010 2011-2015 2016-2020

What does Investment Research entail?

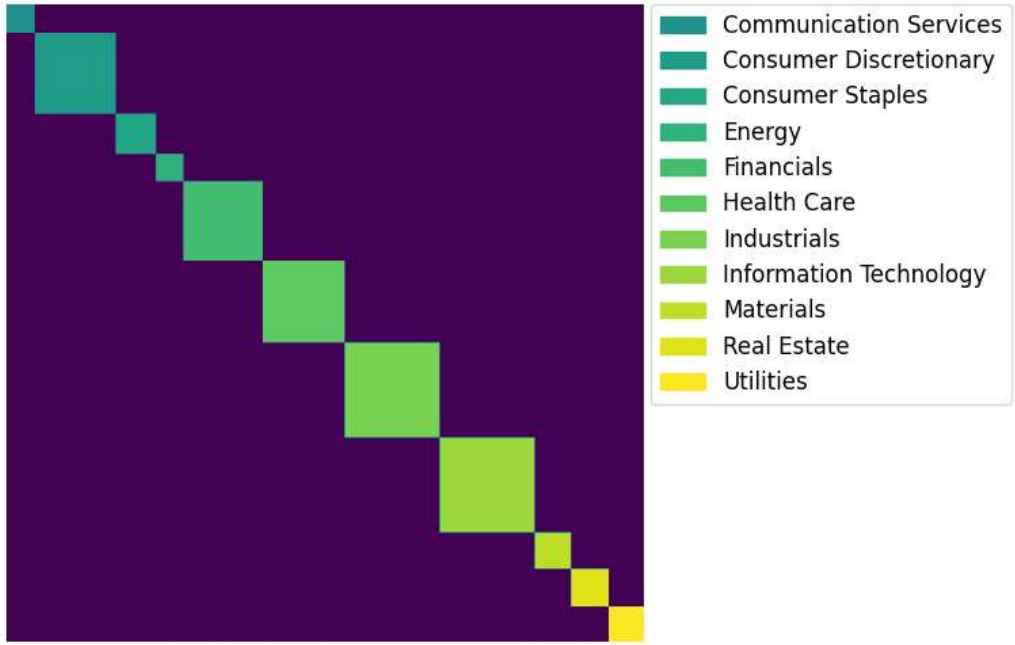
- To identify unique investment opportunities
  - Understanding individual companies' business models
  - Following sector trends and industry dynamics: knowledge accumulation
  - Financial reports are a primary source of information
- Some specific questions in Investment Research:
  - What ***factors*** affect a company's financial performance?
  - How do these ***factors*** change over time? Are there any recognizable ***patterns***?
  - Which are the most ***closely related*** and comparable peers of a given company?
  - Given certain ***macro-level events***, which companies are most likely to be affected?

5 Use Cases

- Identifying Comparable Companies
  - Pairwise cosine similarity matrix of S&P 500 companies based on financial reports in last 10 years
  - Companies arranged by sector classification: lighter-colored regions along the diagonal
  - Validation that our model correctly identifies semantic similarity amongst peers in the same sector
  - Interesting observation: some sector companies exhibit stronger similarity signal than others



(a) Similarity Matrix



(b) Grouping by Sector

What does Investment Research entail?

- To identify unique investment opportunities
  - Understanding individual companies' business models
  - Following sector trends and industry dynamics: knowledge accumulation
  - Financial reports are a primary source of information
- Some specific questions in Investment Research:
  - What ***factors*** affect a company's financial performance?
  - How do these ***factors*** change over time? Are there any recognizable ***patterns***?
  - Which are the most ***closely related*** and comparable peers of a given company?
  - Given certain ***macro-level events***, which companies are most likely to be affected?

5 Use Cases

- Keyword Search for the Most Relevant Companies

Input: *keywords*

1. First, convert keyword into word embeddings and feed into the clustering module to find the most relevant set of semantic groups
2. Next, create a query to traverse the graph and collect all the companies that are connected to the identified semantic groups
3. Rank companies by aggregating the total number of links to the set of semantic groups

Output: *List of Companies*

Keyword	HC Cluster #	SOM Cluster #	Sample Noun Phrases in the Same SOM Cluster
wheat	347	2435	*wheat Thins, increase wheat, wheat flour, wheat milling operation, Gold Medal flour, wheat derivative*
corn	347	2377	*corn acre, healthy corn oil, corn seed sector, herbicide tolerant corn, improved flour*
soybean	347	2435	*poor crop, oilseed, limited soybean, nutrient, livestock feed, aminopyralid insecticide*
meat	348	2203	*hog, fresh pork, frozen pizza, Habit Burger Grill, cattle, taco, meat ingredient, chicken, swine herd,*

Companies most relevant to input search terms:

1. ***General Mills:*** manufacturer and marketer of branded consumer foods
2. ***Tyson Foods*****:** meat processor and marketer
3. ***Campbell Soup*****:** processed food and snack company
4. ***Zoetis:*** a global animal health company
5. ***Archer-Daniels-Midland:*** food processing and commodities trading corporation

Conclusion

Objective: Create a **text mining** tool to aid **investment research**

- ☑ Extract **causal factors** from financial reports
- ☑ Design a **data model** that facilitates data exploration and knowledge discovery
- ☑ Implement an **end-to-end pipeline** that provides a **proof-of-concept** demonstration of solving real-life problems

Future Work:

- ➢ To expand size and scope of the corpus: additional companies, longer timeframe, equity research reports, sector analysis, financial news, etc.
- ➢ To establish a comprehensive ground truth baseline: causality extraction and clustering
- ➢ Web-based user interface with better data visualization; field usability studies

Questions?

Thank you!

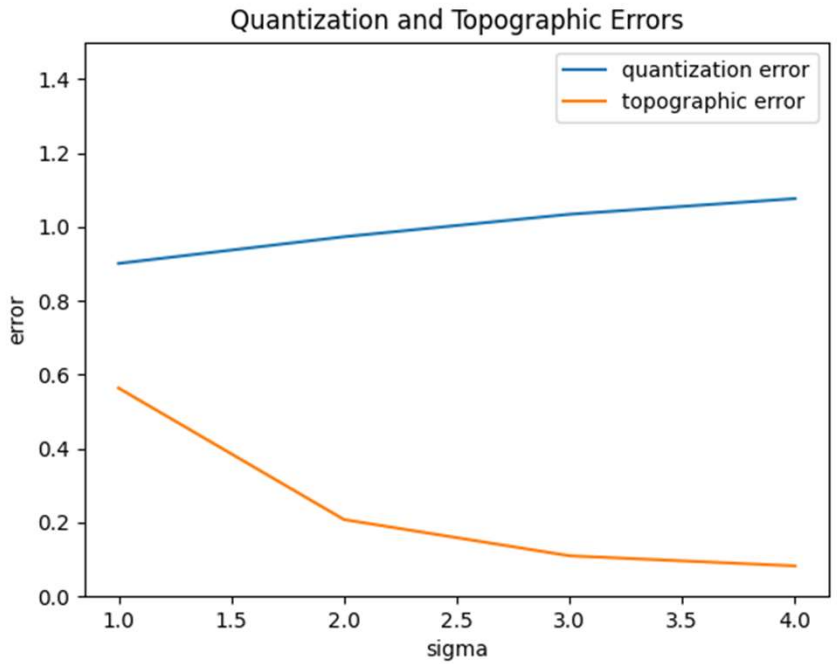
Reference

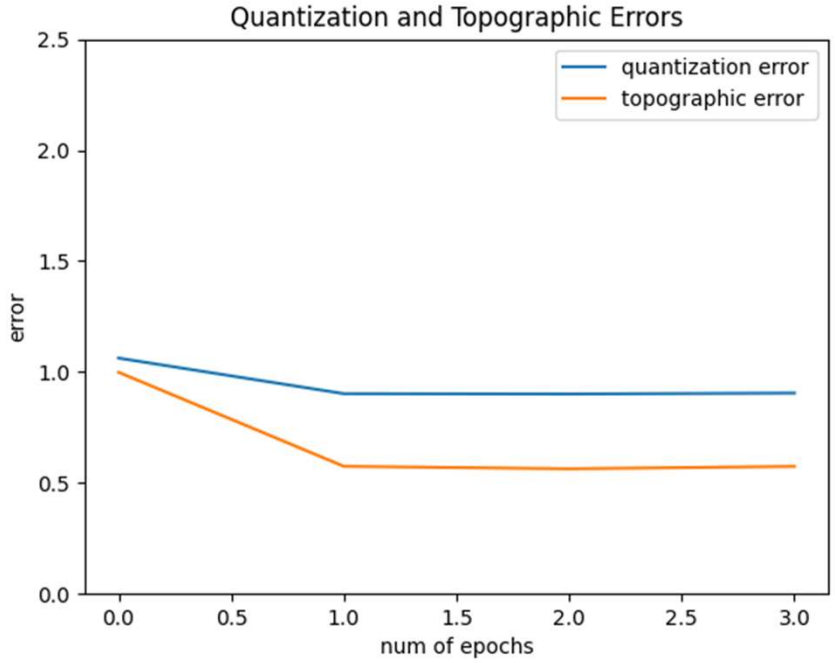
References are the same as the Thesis report

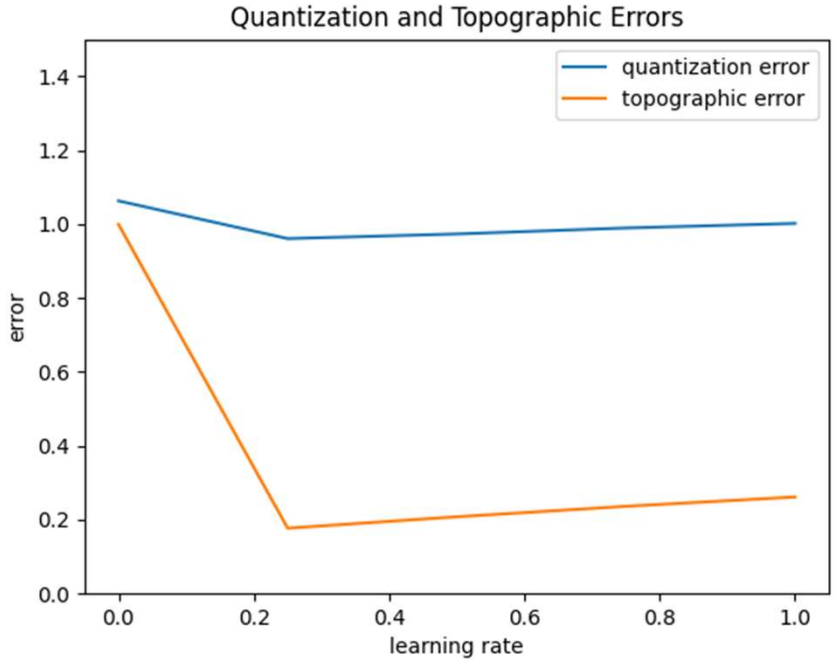
Backup Slides

SOM Parameters

- No. of Neurons: 5x$\sqrt{N}$
- Topology: rectangular 58x58
- Neighbourhood: Gaussian
- Sigma: 2.0
- Learning rate: 0.25
- No. of Epochs: 1







Quantization Error (QE)

- Basic quality measure of SOMs
- Average distance between data points and their assigned nodes
- Smaller values indicate better fit

$$QE(M) = \frac{1}{n}\sum_{i=1}^{n} ||\phi(x_i) - x_i||$$

- where n is the number of data points in the training data and ϕ : D → M is the mapping from the input space D to the SOM M

Topological Error (TE)

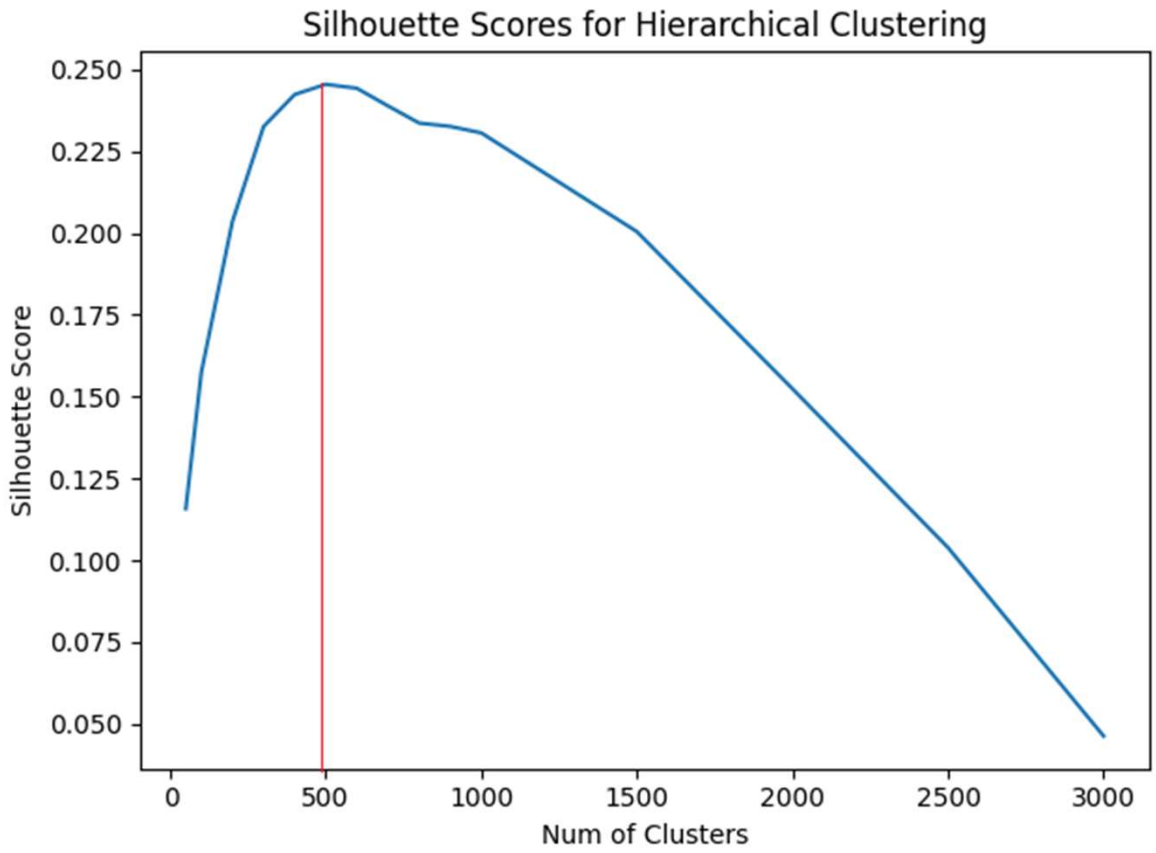
- How well the structure of the input space is modeled by the map
- Find the best-matching and second-best-matching neurons in the map for each input and then evaluating their positions
  - If the two nodes are next to each other, then topology is deemed to have been preserved for this input
  - If not, then this is counted as an error
  - The total number of errors divided by the total number of data points gives the topographic error of the map

$$TE(M) = \frac{1}{n}\sum_{i=1}^{n} t(x_i)$$

- where t(x) = 0 if the best-matching and second-best-matching neurons are neighbors; else t(x) = 1

Silhouette Score

- Mean Silhouette Coefficient of all samples
- Silhouette Coefficient for a sample is (b - a) / max(a, b), where
  - a = mean intra-cluster distance
  - b = mean nearest-cluster distance, i.e., the distance between a sample and the nearest cluster that the sample is not a part of
- The best value is 1 and the worst value is -1
- Values near 0 indicate overlapping clusters
- Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar

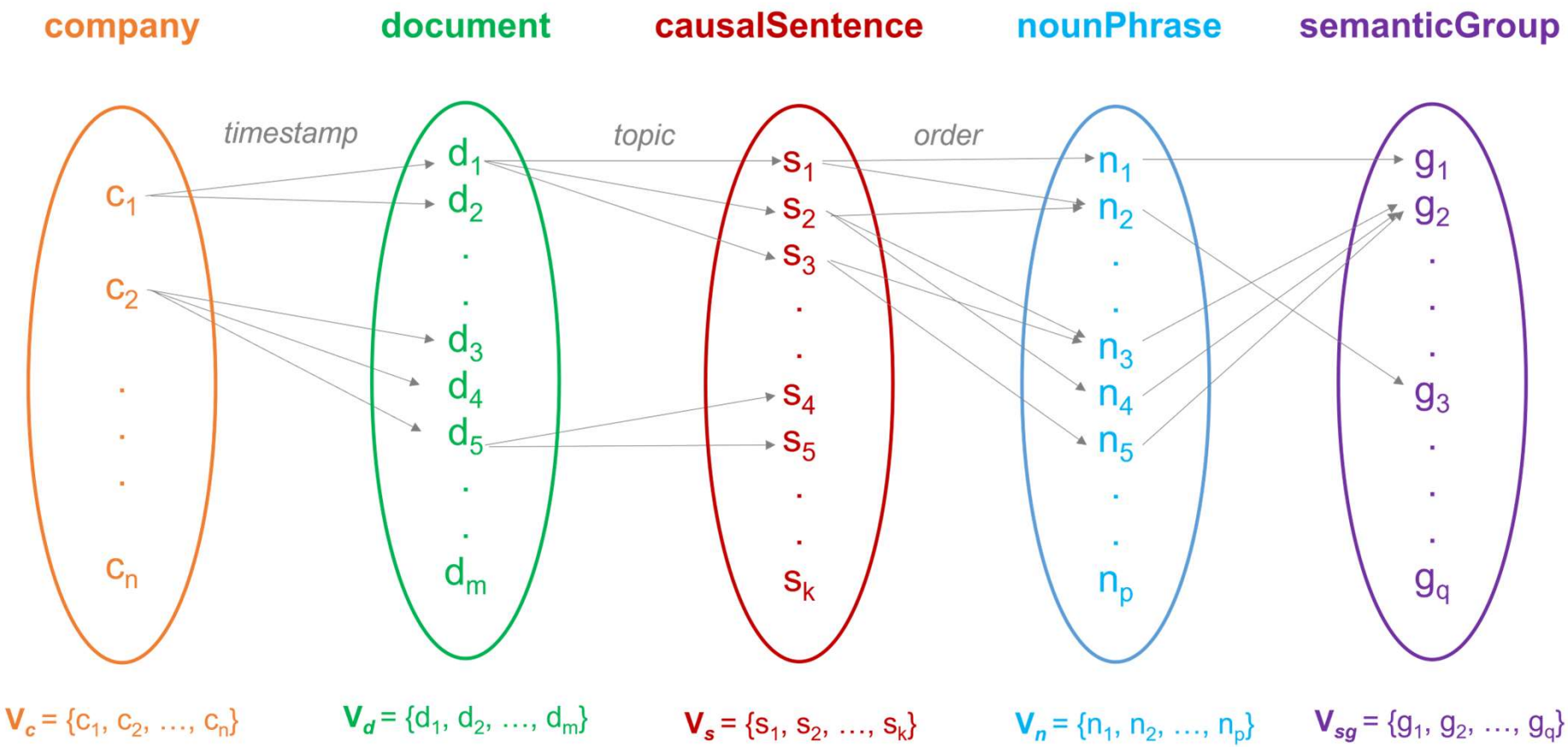


1 Raw Data

Metadata Table

	ticker	cik	name	exchange	10K_files	10Q_files	k_count	q_count	total	GICS Sector
0	MSFT	789019	MICROSOFT CORP	Nasdaq	['0001564590-21-039151', '0001564590-20-034944...	['0001564590-21-051992', '0001564590-21-020891...	21	62	83	Information Technology
1	AAPL	320193	Apple Inc.	Nasdaq	['0000320193-21-000105', '0000320193-20-000096...	['0000320193-21-000065', '0000320193-21-000056...	21	64	85	Information Technology
2	GOOG	1652044	Alphabet Inc.	Nasdaq	['0001652044-21-000010', '0001652044-20-000008...	['0001652044-21-000057', '0001652044-21-000047...	6	19	25	Communication Services
3	AMZN	1018724	AMAZON COM INC	Nasdaq	['0001018724-21-000004', '0001018724-20-000004...	['0001018724-21-000028', '0001018724-21-000020...	18	64	82	Consumer Discretionary
4	TSLA	1318605	Tesla, Inc.	Nasdaq	['0001564590-21-004599', '0001564590-20-004475...	['0000950170-21-002253', '0000950170-21-000524...	11	35	46	Consumer Discretionary

4 Data Model


company
document
causalSentence
nounPhrase
semanticGroup
timestamp
topic
order
c_1
c_2
c_n
d_1
d_2
d_3
d_4
d_5
d_m
s_1
s_2
s_3
s_4
s_5
s_k
n_1
n_2
n_3
n_4
n_5
n_p
g_1
g_2
g_3
g_q
$V_c = \{c_1, c_2, \ldots, c_n\}$
$V_d = \{d_1, d_2, \ldots, d_m\}$
$V_s = \{s_1, s_2, \ldots, s_k\}$
$V_n = \{n_1, n_2, \ldots, n_p\}$
$V_{sg} = \{g_1, g_2, \ldots, g_q\}$

Three Real-life Observations

1. Financial reports as a primary source of information
   - Experienced analyst knows where to look for relevant information
   - Credit attribution in Management Discussion and Analysis (MD&A) sections
2. Association, abstraction and categorization
   - Naturally associate similar causal factors into meaningful groups
   - Knowledge accumulation and pattern recognition
3. Dynamic nature of company-factor relations
   - One-off, recurring, cyclical
   - Evolution with time

5 Use Cases

- ## Identifying Comparable Companies

The top five most similar companies for JP Morgan Chase:

1. Regions Financial (cosine similarity = 0.728)
2. Hartford Financial Services Group (cosine similarity = 0.713)
3. Bank of America (cosine similarity = 0.698)
4. *Iron Mountain (cosine similarity = 0.677)*
5. Willis Towers Watson (cosine similarity = 0.653)

Apart from *Iron Mountain*, which is a secure data backup and storage company (often used by financial institutions), all of the other top-ranking companies are in the financial sector, as expected.

5 Use Cases

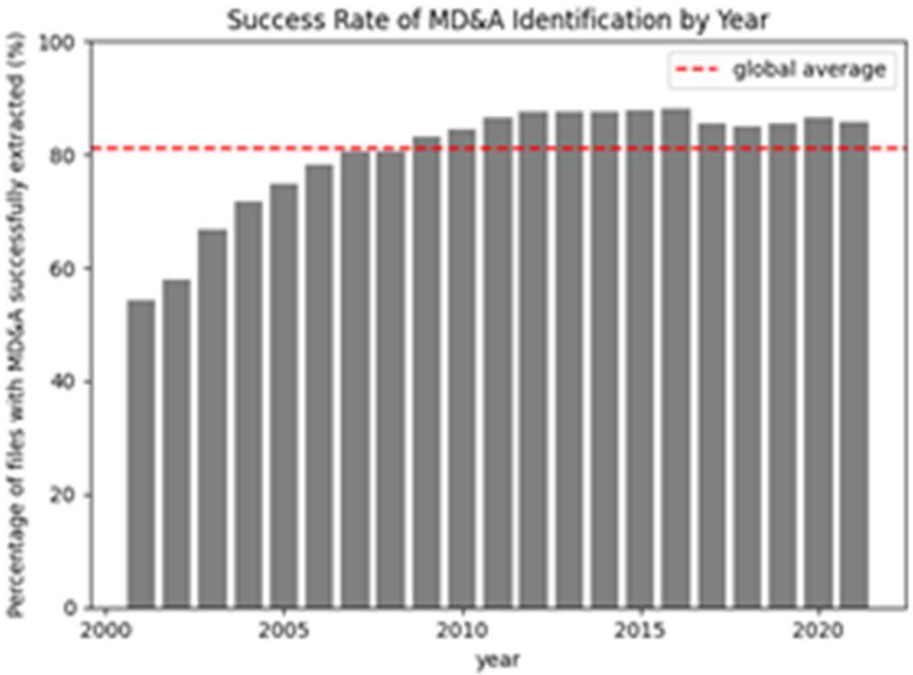
- Keyword Search for the Most Relevant Companies

Keywords	Top 5 Most Relevant Companies
aircrafts, flights, fuel price	Alaska Air Group, Southern, American Airlines Group, United Parcel Service, Southwest Airlines
cloud computing, business analytics, cloud storage, data management	Amazon, Akamai Technologies, CF Industries Holdings, Northrop Grumman, Equinix
global supply chain disruption, inflation, pandemic, Covid-19	Prudential Financial, Metlife, Sysco, United Parcel Service, Cognizant Technology Solutions

Note that the presentation is at most 25 minutes, followed by a Q&A session. Please also keep in mind that there needs to be a clear **problem statement**, **objectives**, and the methods should be presented in a way that people can understand what you have done (what the **challenges** and **contributions** are) without having read the thesis.

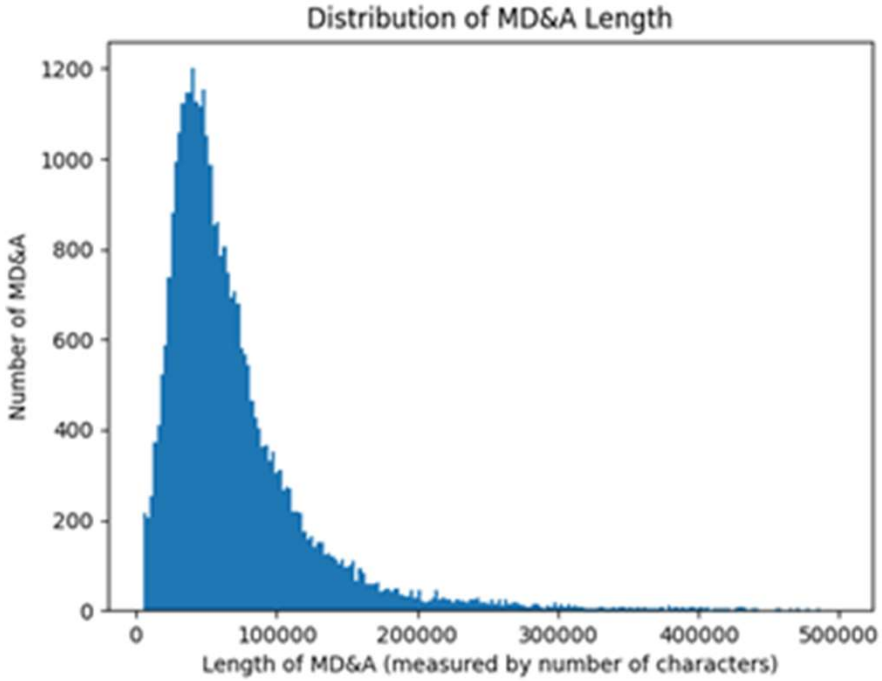
1 Raw Data

Success rate of MD&A Extraction



MD&As are extracted from c. 81% of all downloaded files

Distribution of MD&A Length



The average length of MD&A is c. 70,217 characters

Text Mining Taxonomy