An Algorithmic Framework for Systematic Literature Reviews

A Case Study for Financial Narratives

A Quantitative Finance Perspective

Gabin Taibi Joerg Osterrieder

BFH Bern University of Applied Sciences

March 27, 2026

Notes — Slide 1

Welcome. This talk presents a dual contribution: a reproducible algorithmic framework for systematic literature reviews and its application to the rapidly growing field of financial narratives. We have 45 minutes, so I will go deeper into the methodology and the findings than the short version. Happy to take questions at the end or throughout.

The Narrative Turn in Finance

- **Shiller (2017, 2019):** Economic narratives go viral, drive booms, panics, recessions — "narrative economics" as a research program
- **Grossman & Stiglitz (1980):** Information heterogeneity persists in equilibrium — narratives fill the gap between prices and beliefs
- **Bybee et al. (2023):** News-based narrative factors explain cross-sectional risk premia beyond traditional factor models
- **Bybee et al. (2024):** News attention tracks business cycles, forecasts aggregate stock returns out-of-sample
- **Flynn & Sastry (2024):** Belief-driven macroeconomic fluctuations extracted from firm disclosures

Narratives have measurable, quantifiable effects on financial markets.

Notes — Slide 2

Narratives have measurable, quantifiable effects on markets — not soft science. Shiller's Nobel lecture made the case. Since then, quantitative evidence has accumulated rapidly. Bybee and coauthors show narrative factors rival Fama-French factors in explanatory power.

What IS a Financial Narrative?

- **Somers (1994):** Structured accounts constructed through *emplotment* — events gain meaning via causal ordering
- **Shiller (2017):** Contagious stories that spread virally and drive aggregate economic activity
- **Tuckett & Nikolic (2017):** *Conviction narratives* — belief structures formed under radical uncertainty
- **Roos & Reccius (2024):** Socially shared, action-oriented stories with temporal and causal structure

Working Definition

"Structured, transmissible interpretations concerning financial markets, based on available information, possessing causal temporal structure."

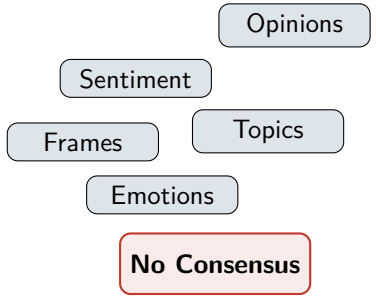
Not just sentiment.
Not just topics.
Not just opinions.

Notes — Slide 3

We need a precise definition to anchor the talk. A financial narrative is not just sentiment — it has structure, causality, and transmissibility. Somers gives us the sociological foundation: emplotment. Shiller adds virality. Tuckett adds conviction under uncertainty. Our working definition synthesizes these views. This distinction matters because most papers reduce narratives to sentiment scores, which discards most of the information.

The Problem — Fragmentation

- "Financial narrative" means different things to different researchers
- Most studies reduce narratives to **sentiment** — a lossy compression
- Methods range from dictionary-based to LLM-based — **no consensus**
- No unified typology, no standard benchmarks
- Hard to build on each other's work, hard to compare results



Notes — Slide 4

The field is growing fast but incoherently. This motivates a systematic review. Everyone uses the term "financial narrative" but they mean very different things. Most default to sentiment, which discards structure, causality, and temporal dynamics.

Research Questions & Contribution

Research Questions

RQ1 How can NLP be used to **quantify and model** financial narratives?

RQ2 How is "financial narrative" **defined and applied** in the literature?

Dual Contribution

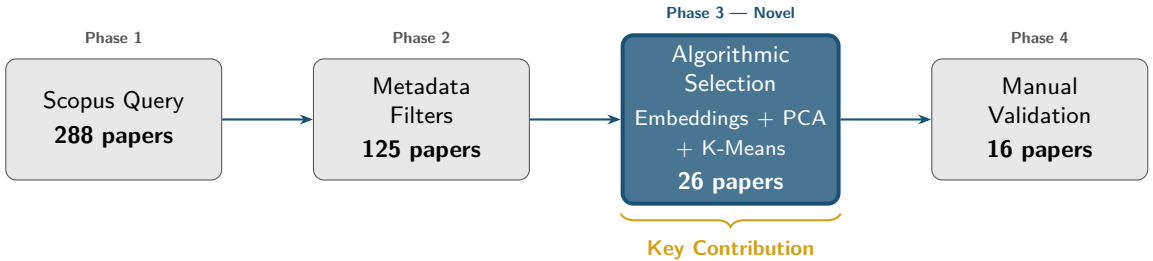
(a) Algorithmic SLR Framework — reproducible, embedding-based paper selection replacing subjective screening

(b) Structured Synthesis — concept matrix revealing gaps, two-stream taxonomy, future directions

Notes — Slide 5

The methodology IS a contribution, not just scaffolding. Most SLRs treat their method as boilerplate. Ours is reusable, quantitative, and addresses the scalability problem of systematic reviews in fast-moving fields.

SLR Pipeline Overview



- **Phase 3** replaces subjective screening with semantic embedding similarity
- 288 → 125 → 26 → 16: systematic, reproducible reduction

Notes — Slide 6

Phase 3 is the novel contribution. Everything else is standard SLR methodology per PRISMA. The key innovation is replacing human judgment in screening with embedding-based semantic similarity and unsupervised clustering. In the next slides I will walk through each phase in detail.

Phase 1 — Query Design

Three Query Blocks (Scopus)

1. **Block A:** "financial narrative" OR "economic narrative" in *Title / Abstract / Keywords*
2. **Block B:** narrative* AND (NLP-terms) AND (financial-terms) in *Title*
3. **Block C:** Same combination as Block B, searched in *Keywords*

Design Rationale

- Broad enough for **recall**: captures heterogeneous terminology
- Narrow enough for **precision**: excludes health, psychology, literary studies
- Boolean OR across blocks, AND within compound terms

Initial Yield

288 papers

from Scopus, before any filtering

Notes — Slide 7

Show the actual query structure. Quants appreciate precise specifications. The three-block design is deliberate: Block A catches papers that explicitly use the term "financial narrative." Blocks B and C catch papers that discuss narrative methods in financial contexts without using the exact phrase. This union maximizes recall while the financial-terms constraint maintains precision.

Phase 2 — Metadata Filtering

Inclusion Criteria

- English language only
- Published 2011 or later
- Journal articles only (peer-reviewed)
- SCImago discipline filtering

Excluded Disciplines

- Medicine & Health Professions
- Arts & Humanities
- Psychology
- Environmental Science
- Other non-financial domains

Retained Disciplines

- Economics, Econometrics & Finance
- Computer Science
- Social Sciences
- Business, Management & Accounting
- Mathematics & Decision Sciences

Result

288 $\rightarrow$ **125** papers

163 removed by metadata criteria

Notes — Slide 8

Standard SLR practice. The SCImago discipline filtering is important because "narrative" is a term used across many fields — health narratives, literary narratives, psychological narratives. Without discipline filtering, more than half the results are irrelevant. The novel part comes in Phase 3.

Phase 3a — Screening Statements

Six Screening Criteria (derived from RQ1 & RQ2)

1. The paper focuses on **financial or economic narratives**
2. The paper studies narratives in the context of **financial markets or macroeconomics**
3. The paper presents **empirical or computational methods**
4. The paper applies **NLP, text mining, or machine learning** methodology
5. The paper uses **financial text data** (news, filings, social media, reports)
6. The paper addresses questions **aligned with RQ1 or RQ2**

GPT-3.5 Paraphrasing

- Each statement paraphrased into multiple variants
- Variants embedded separately, then **averaged**

Zero-Shot Latent Embedding

Define research criteria in natural language

↓

Embed them

↓

Filter by cosine similarity

Notes — Slide 9

You define research criteria in natural language, embed them, and cosine similarity does the filtering. This is a zero-shot approach — no training data needed. The GPT-3.5 paraphrasing step is a robustness measure: instead of relying on the exact phrasing of a single statement, we generate multiple variants and average their embeddings. This smooths out the sensitivity to any particular wording choice.

Phase 3b — Embedding & Similarity

Embedding Model

- `text-embedding-3-small` (OpenAI)
- Unit-normalized vectors
- Cosine similarity = dot product
- 1536-dimensional space

Similarity Matrix

$$\mathbf{S} \in \mathbb{R}^{125 \times 6}$$

$$S_{ij} = \cos(\mathbf{e}_i^{\text{paper}},\ \mathbf{e}_j^{\text{criterion}})$$

125 papers $\times$ 6 screening criteria

Paper Representation

- Each paper: **title + abstract + keywords**
- Concatenated into single text
- Single embedding per paper

Standardization

- Z-score normalization per column
- Ensures criteria contribute equally
- Removes scale differences

Notes — Slide 10

We measure how semantically close each paper is to our ideal criteria. The similarity matrix is the key data structure. Each row is a paper, each column is a screening criterion, and the entry is the cosine similarity between the paper embedding and the criterion embedding. Z-score normalization ensures no single criterion dominates simply because of higher baseline similarity.

Phase 3c — Dimensionality Reduction & Clustering

PCA Diagnostics

- **KMO** = 0.815 (meritorious sampling adequacy)
- **Bartlett's test:** $p < 0.001$
- **Condition number** = 370
- **4 components** at 99% variance explained
- Reduces 6 → 4 dimensions

Why PCA?

Screening statements are correlated.
PCA extracts orthogonal relevance dimensions.

K-Means Clustering ($k = 3$)

Cluster	Size	Avg Rel.
High relevance	26	0.507
Medium relevance	*excluded*	
Low relevance	*excluded*	

Selection Decision

- Silhouette score: **0.352**
- Medium cluster excluded after boundary inspection
- Conservative: prefer precision over recall
- → **26 papers** selected

Notes — Slide 11

KMO and Silhouette are diagnostics quants already know. The KMO of 0.815 means the correlation structure is suitable for factor analysis. The Silhouette score of 0.352 indicates moderate but clear cluster separation. We chose to exclude the medium cluster conservatively — better to miss a borderline paper than include an irrelevant one. The high-relevance cluster of 26 papers has a mean relevance of 0.507, well separated from the medium cluster.

Stability & Robustness

PCA Variance Thresholds

- Tested 80%, 85%, 90%, 95%, 98% explained variance
- **All thresholds yield the same 26 papers**
- Result insensitive to PCA hyperparameter

Conclusion

Results robust to methodological choices

Framework produces stable outputs across reasonable parameter ranges

Clustering Algorithm Comparison

Method	Papers	Avg Rel.	Silhouette
K-Means	26	0.504	0.347
GMM	fewer	higher	worse cov.
Agglom.	50	—	highest

- K-Means: best balance of size, relevance, and separation
- GMM: fewer papers, higher relevance, worse coverage
- Agglomerative: 50 papers, highest Silhouette, diluted relevance

Notes — Slide 12

The robustness check quants demand. Two axes of variation tested: PCA variance threshold and clustering algorithm. On the PCA side, all thresholds from 80 to 98 percent yield the identical 26-paper set. On the clustering side, K-Means provides the best tradeoff: balanced cluster size, reasonable Silhouette, and sufficient coverage. GMM is too selective, Agglomerative too permissive.

Phase 4 — Manual Validation

Exclusion Reasons

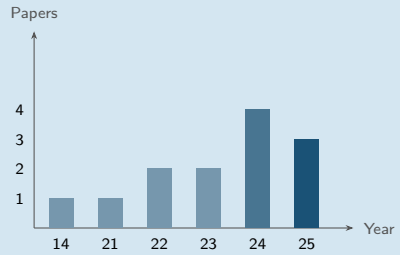
- 2 papers **inaccessible** (paywalled, no institutional access)
- 4 papers were **workshop proceedings** (not peer-reviewed journals)
- Remaining exclusions: insufficient methodological depth, tangential focus

Final Corpus

26 → **16** papers

10 removed by manual review

Temporal Distribution



Majority of papers published 2022–2025. Field is accelerating.

Notes — Slide 13

Quick slide. The manual validation step is small and justified. Only 10 papers were removed, and the reasons are transparent: inaccessibility, non-peer-reviewed venue, or insufficient methodological depth. The temporal distribution confirms the field is recent and accelerating, with the majority of papers from the last three years.

Two Streams of Literature

Notes — Slide 14

Understanding = "what is a narrative?" Modeling = "what can narratives predict?" This two-stream taxonomy emerged from the concept matrix. Understanding papers build the linguistic and conceptual infrastructure. Modeling papers build the predictive applications. I will now walk through each stream in more detail than the short version.

Understanding — Annotation & Corpora

Hu et al. (2021)

- Annotation scheme integrating **intra-sentential labeling** with economic terminology
- Corpus: central bank speeches, 10-K filings, news articles, social media
- One of the first **structured annotation frameworks** for financial narratives

Zmandar et al. (2022)

- **CoFiF Plus** — first large-scale French financial narrative corpus
- Multi-task annotations for classification, summarization, NER
- Narratives framed as **communicative devices** for shaping investor perceptions

Why This Matters

Without annotation and corpora, downstream modeling has no foundation.
These papers build the **data layer** for the entire field.

Notes — Slide 15

Data layer papers. Without annotation and corpora, downstream modeling has no foundation. Hu et al. is important because it goes beyond document-level labels to intra-sentential annotation — marking which parts of a sentence carry narrative content. Zmandar et al. extends the field beyond English, which is critical for any claim of generalizability. Both papers treat narratives as more than sentiment: they are communicative devices with internal structure.

Understanding — Structure & Semantics

- **Sy et al. (2023):** BERT ensembles for **argument mining** in earnings calls
  - Identifies claims, premises, and argumentative relations
  - Moves beyond sentiment to **logic and discourse coherence**
- **Liu et al. (2024):** *Financial-STS* — sentence-level semantic textual similarity
  - Detects **year-over-year semantic drift** in corporate disclosures
  - Measures how narratives evolve across reporting periods
- **Roos & Reccius (2024):** Formal definition of financial narratives
  - Clear distinction: **narratives** $\neq$ **topics** $\neq$ **sentiment**
  - Narratives are socially shared, action-oriented stories
  - Most rigorous conceptual contribution in the corpus

Roos & Reccius draws a clear line between narrative and sentiment — a distinction most modeling papers ignore.

Notes — Slide 16

Roos and Reccius is the conceptual anchor of the understanding stream. They show that a narrative is not just a positive or negative signal — it is a structured, causal, temporally ordered account that is socially shared and drives action. Sy et al. operationalizes this by identifying argumentative structure in earnings calls. Liu et al. provides a tool for tracking how narratives change over time. Together, these papers build the infrastructure that the modeling papers mostly lack.

Modeling — Early & Historical

Tuckett et al. (2014)

- **Conviction narratives**: "phantastic objects" that sustain belief under radical uncertainty
- Directed sentiment from Reuters and Enron corpora
- Regime detection: narrative shifts precede market regime changes
- Earliest paper in the corpus

Hsu et al. (2021)

- Historical Chinese news from the 1930s
- Silver Purchase Act: policy narrative drove capital flight
- **Textual frequency + regularized regression**
- Demonstrates narrative effects in a clean historical laboratory

Roots going back a decade — this is not just LLM hype.
Narrative modeling predates the transformer revolution.

Notes — Slide 17

Important to show historical depth. Tuckett et al. from 2014 is the earliest paper in our corpus and introduces the concept of conviction narratives — stories that sustain action even when evidence is ambiguous. Hsu et al. uses historical Chinese news to study how policy narratives during the Silver Purchase Act drove capital outflows. Both demonstrate that narrative effects are robust and predate modern NLP. This is not just LLM hype — it is a decade-long research program.

Modeling — Predictive Power

Returns & Volatility

- **Chen et al. (2022):** COVID narrative virality → extreme realized volatility across markets
- **Borup et al. (2023):** Narrative investor expectations **outperform** sentiment indices for return prediction

Inflation Forecasting

- **Hong et al. (2025):** 880K WSJ articles → narrative factors **outperform macro models** for inflation forecasting
- Text beats macroeconomic data

Sector & Market Returns

- **Ma et al. (2024):** Narrative Energy Index (NEI) predicts both sector and aggregate market returns

Key Takeaway

Four different targets.
Four different methods.
Consistent outperformance.

Each study represents a potential **alpha signal** or **risk indicator**: returns, volatility, inflation, sector rotation

Notes — Slide 18

MONEY SLIDE. Four targets, four methods, consistent outperformance over traditional approaches. Let the audience absorb this. Chen shows volatility transmission via narrative virality. Borup shows narrative expectations beat sentiment. Ma constructs a narrative energy index for sector rotation. Hong shows 880 thousand WSJ articles outperform macroeconomic models for inflation. This is the slide that justifies the entire field for a quant audience.

Modeling — Emerging Approaches

- **Miori & Petrov (2023):** GPT + graph theory
  - Narrative network construction from GPT-extracted entities and relations
  - **Narrative fragmentation** as a proxy for systemic risk
- **Stander (2024):** FinBERT sentiment for South African news
  - Credit risk modeling under IFRS 9 framework
  - Extends narrative finance beyond US/European markets
- **Agarwal et al. (2024):** Emotion metrics in financial media
  - Fear, anger, surprise $\rightarrow$ returns, volume, volatility
  - Particularly powerful during **market bubbles**
- **Taffler et al. (2024):** Crisis emotion dictionaries
  - Custom lexicons for crisis contexts
  - Narrative affect explains $>50\%$ of returns **in extreme periods**

Notes — Slide 19

GPT-based networks and crisis emotion modeling represent the frontier. Miori and Petrov use GPT to extract entities and relations, then build narrative networks — fragmentation of these networks signals systemic risk. Stander extends the field geographically to South Africa and methodologically to credit risk. Agarwal et al. show that granular emotions, not just positive/negative sentiment, matter for asset prices. Taffler et al. demonstrate that during crises, narrative affect explains more than half of return variation — an extraordinary number.

Concept Matrix

Paper	Theory			Context				Method			
	Sentiment	*Structured*	*Formal Def.*	*Equity*	*Macro*	*Crisis*	*Other*	*Lexicon*	*Topic Model*	*Transformer*	*LLM*
Tuckett et al. (2014)	■						■	■			
Hu et al. (2021)	■	■			■						
Chen et al. (2022)	■			■		■			■		
Zmandar et al. (2022)		■					■		■	■	
Borup et al. (2023)	■			■	■				■		
Sy et al. (2023)		■		■						■	
Miori & Petrov (2023)		■				■					■
Liu et al. (2024)		■		■						■	
Ma et al. (2024)	■			■	■				■		
Roos & Reccius (2024)		■	■				■				
Hong et al. (2025)					■				■	■	
Coverage	*5/11*	*6/11*	*1/11*	*5/11*	*4/11*	*2/11*	*3/11*	*1/11*	*4/11*	*4/11*	*1/11*

■ = covered ■ = **gap / opportunity**

Notes — Slide 20

Take a moment with this table. The power is in what is MISSING — the white space is the opportunity. Only 1 paper provides a formal definition. Only 1 uses LLMs. Only 2 study crisis contexts. The intersection of "formal definition + transformer + equity" is completely empty. I will now unpack the patterns we see in the Theory, Context, and Method dimensions separately.

TCM Analysis — Theory Dimension

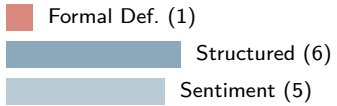
What We See

- Most papers: **implicit or semi-structured** conceptualization of narrative
- Dominant proxy: sentiment scores, emotion distributions, topic frequencies
- Few papers use **structured representations**: annotation, argument mining, graph-based
- Formal definitions **rare** — mainly Roos & Reccius (2024)

The Implication

Narratives are treated as **inputs to measure**, not **objects to model**.

The field lacks a shared formal apparatus.



Notes — Slide 21

Narratives treated as inputs to measure, not objects to model. This is the core theoretical gap. Most papers grab text, compute some numerical proxy — sentiment, topic weights, emotion scores — and feed it into a regression. Very few ask: what IS the object we are measuring? How should it be represented formally? This is like doing portfolio optimization without a formal definition of return and risk. Roos and Reccius is the exception, but their work is purely conceptual — it has not yet been operationalized computationally.

TCM Analysis — Context & Method

Context Concentration

- **Equity and macro dominate**: 5/11 each
- Crisis context: only 2/11
- Heavy reliance on **news and filings**
- **Absent**: fixed income, FX, derivatives, commodities, social media (as primary source)

Method Mismatch

- **Modeling papers**: lexicon + LDA dominate
- **Understanding papers**: contextual embeddings (BERT, transformers)
- LLMs: only Miori & Petrov (1/11)
- Financial side: regressions and basic Granger causality
- **No structural or joint models**

Gap

Most of the financial universe is unstudied through the narrative lens.

Key Insight

The **best NLP methods** appear in understanding papers that do not model markets. The **market-facing papers** use older NLP.

Notes — Slide 22

Method-context mismatch is the key insight. The papers that use the most sophisticated NLP — transformers, argument mining, semantic similarity — are the understanding papers that do not model financial outcomes. The papers that actually predict returns and volatility use older methods: lexicons, LDA, basic regressions. Nobody has combined state-of-the-art NLP with state-of-the-art financial modeling. This is Gap 1 in a nutshell.

Four Research Gaps

G1 No integration of structured representations + semantic methods + financial modeling

Papers either define narratives rigorously OR model them quantitatively — never both

G2 Contextual embeddings largely absent from financial narrative applications

Most modeling papers use topic models or lexicons; transformers appear only in understanding papers

G3 Empirical scope is narrow

Few asset classes beyond equities, limited data sources beyond news and filings

G4 Narratives treated as exogenous, never as endogenous market components

No paper models how narratives form, spread, compete, and reshape markets

Notes — Slide 23

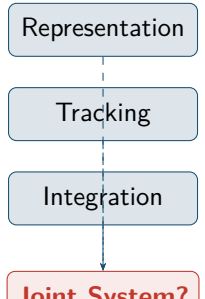
Frame each gap as an opportunity for the audience. G1 is the biggest: nobody has combined all three elements. G2 is the most immediately actionable. G3 is the broadest. G4 is the most ambitious: treating narratives as endogenous requires agent-based or general-equilibrium modeling. Each gap is a viable research program, and I will now unpack G1 in more detail.

Gap Deep Dive — The Missing Integration

Gap 1 Unpacked

What does not exist yet: a system that...

1. **Represents** narratives structurally (causal chains, temporal ordering, agents)
2. **Tracks** them with contextual embeddings (not bag-of-words proxies)
3. **Integrates** into financial models — not as an input feature, but as a **model component**



Quant Analogy

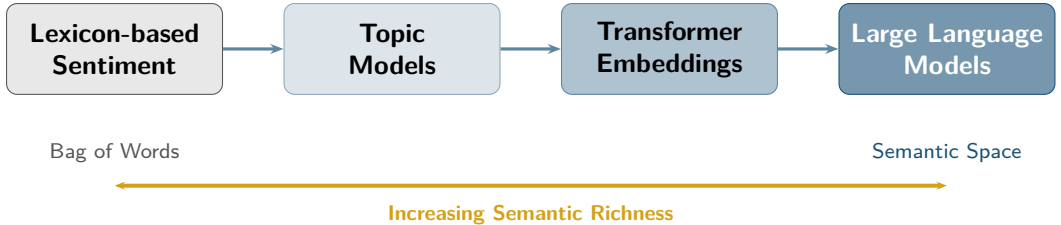
Like having factor construction, risk modeling, and portfolio optimization as **separate disconnected systems**.

Nobody has built the **integrated pipeline**.

Notes — Slide 24

Make the gap concrete and relatable. The analogy to quantitative finance is deliberate: imagine doing factor construction, risk modeling, and portfolio optimization in three separate disconnected systems. That is where narrative finance is today. The pieces exist — structured representations in understanding papers, contextual embeddings in a few NLP papers, market modeling in the prediction papers — but nobody has connected them into an integrated system. This is the single biggest opportunity in the field.

From Sentiment to Semantic Space



Key Insight

Narratives $\approx$ **trajectories in semantic embedding space**

Position · Velocity · Clustering · Divergence

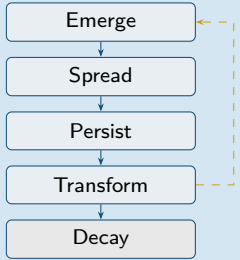
Familiar mathematical framework, fundamentally new input signal

Notes — Slide 25

Trajectories in embedding space are analogous to trajectories in factor space. Familiar math, new input. A narrative is not a number — it is a path through a high-dimensional space. Its position tells you what is being discussed. Its velocity tells you how fast attention is shifting. Its clustering tells you consensus vs. fragmentation. Quants already know how to work with these objects.

Narratives as Dynamic Systems

Narrative Life Cycle



Embedding-Space View

- **Position** = narrative state (what is being said)
- **Velocity** = rate of change (how fast attention shifts)
- **Clustering** = coalescence (consensus forming)
- **Divergence** = fragmentation (competing narratives)

Future Directions

- Latent intensity scores for narrative strength
- New emergence detection via novelty metrics

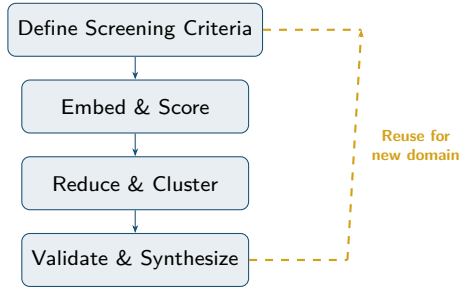
Notes — Slide 26

Plant the seed that narrative dynamics can use math the audience already knows. Stochastic processes, state-space models, point processes — all are candidates for modeling narrative dynamics. The life cycle view is intuitive: narratives emerge, spread virally, persist, transform into new narratives, and eventually decay. In embedding space, each phase has a mathematical signature. Emergence is a new cluster forming. Spread is the cluster growing. Coalescence is multiple clusters merging. Fragmentation is a cluster splitting. The open question is whether this can be formalized into a tractable model — I believe it can.

A Reusable SLR Framework

Design Principles

- Works for **any fast-moving research domain**
- Semantic relevance assessment, not lexical matching
- Quantitative quality metrics replace subjective judgment
- **Reproducible**: same inputs → same outputs
- **Scalable**: handles hundreds to thousands of papers



Applicability

Climate finance · RegTech · Crypto assets · ESG · Any domain with rapid publication growth

Notes — Slide 27

Brief — secondary contribution but important for anyone doing literature surveys. The framework is domain-agnostic. Swap out the screening criteria and it works for climate finance, RegTech, crypto, or any other fast-growing field. The key advantage over manual screening is reproducibility and scalability.

Practical Implications by Role

Signal Builders

- Upgrade from lexicon-based sentiment to **embedding-based narrative extraction**
- Richer signal: structure, causality, intensity — not just polarity
- Demonstrated alpha in returns, volatility, inflation

Macro Forecasters

- Narrative inflation and cycle indicators are **competitive with standard macro models**
- 880K articles outperform Survey of Professional Forecasters
- Complementary signal, not substitute

Risk Managers

- Monitor **narrative fragmentation** as an early indicator of systemic stress
- Track **crisis emotions** (fear, anger) for tail risk signals
- Narrative affect explains $>50\%$ of returns in

Researchers

- Algorithmic SLR framework is **reusable and domain-agnostic**
- Concept matrix identifies concrete gaps
- Four actionable research programs ready for proposal

Notes — Slide 28

Translate findings into actionable takeaways for each audience segment. Signal builders get a concrete upgrade path: move from lexicons to embeddings. Risk managers get two new monitoring channels: narrative fragmentation and crisis emotions. Macro forecasters get a competitive alternative to standard models. Researchers get both a reusable methodology and a map of unexplored territory. Every person in this room should find something directly relevant.

Limitations

- **Single database:** Scopus only — may miss preprints (arXiv, SSRN) and grey literature
- **Abstract-level screening:** Full-text analysis might recover papers whose abstracts understate relevance
- **Closed-source embedding model:** OpenAI `text-embedding-3-small` — reproducibility depends on API stability
- **Open-source vs. closed-source tradeoff:** Open models (e.g., Sentence-BERT) offer reproducibility but may sacrifice embedding quality
- **Medium-relevance cluster excluded:** Conservative choice; some relevant papers may sit at the cluster boundary
- **Scalability:** Embedding API costs grow linearly; full-text embedding of thousands of papers requires cost management
- **Temporal snapshot:** Literature evolving rapidly — results reflect the state at time of query

Notes — Slide 29

Quants respect transparency about limitations. Every one of these is fixable. Multi-database queries are straightforward. Full-text analysis requires more compute but no methodological change. Open-source embeddings from Hugging Face can replace the OpenAI model, though with a potential quality tradeoff. The scalability concern is real but manageable — embedding costs are falling rapidly. We chose pragmatism over perfection for the first iteration.

Future Directions & Call to Action

1. **Unified narrative typologies**
   Agree on what a "financial narrative" is — operationalize Roos & Reccius (2024)
2. **Cross-market, cross-asset studies**
   Beyond equities + news: commodities, fixed income, derivatives, social media, earnings calls
3. **Joint narrative–market models**
   Link narrative measures to quantitative market data in unified econometric or ML frameworks
4. **Open benchmarks for narrative extraction**
   Shared datasets, annotation standards, leaderboards for financial NLP
5. **Narratives as endogenous**
   Model formation, evolution, competition, and market impact jointly — the frontier

Notes — Slide 30

These are specific, actionable research programs. Each one could be a PhD thesis or a multi-year project. Item 5 is the most ambitious and the most important: treating narratives not as exogenous inputs but as endogenous components of the market system. This requires new modeling paradigms — agent-based, network-based, or general-equilibrium approaches.

Summary

What We Built

- **Framework:** Reproducible, scalable, semantic SLR pipeline
- **Application:** Structured synthesis of financial narrative literature

What We Found

- Field advancing rapidly but **fragmented**
- Two streams: **understanding** (what?) vs. **modeling** (what predicts?)
- Four clear research gaps with actionable directions

Key Message

Narratives are **central** to financial markets, not peripheral.

Thank You!

Gabin Taibi
gabin.taibi@bfh.ch

Joerg Osterrieder
joerg.osterrieder@bfh.ch

Paper available upon request

Notes — Slide 31

Crisp summary. Invite questions. The paper is available and we welcome collaboration on any of the future directions discussed. Thank you for your attention.

Backup: Key References

Foundational

- Shiller, R. (2017). Narrative Economics. *AER*.
- Shiller, R. (2019). *Narrative Economics*. Princeton UP.
- Grossman, S. & Stiglitz, J. (1980). On the Impossibility of Informationally Efficient Markets. *AER*.

Narrative Factors

- Bybee, L. et al. (2023). Business News and Business Cycles. *JF*.
- Bybee, L. et al. (2024). Narrative Asset Pricing. *JFE*.
- Borup, D. et al. (2023). The Anatomy of Sentiment-Driven Fluctuations. *JFQA*.

Modeling & Applications

- Hong, T. et al. (2025). News and Inflation Expectations. *WP*.
- Miori, L. & Petrov, S. (2023). GPT-Based Narrative Risk. *WP*.
- Ma, Y. et al. (2024). Narrative Energy Index. *IRFA*.
- Flynn, J. & Sastry, K. (2024). Belief-Driven Fluctuations. *WP*.

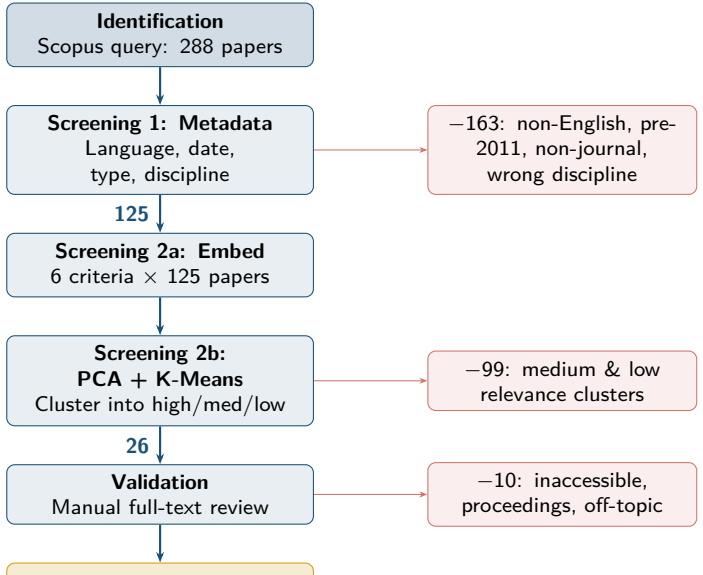
Definitions & Methodology

- Roos, M. & Reccius, M. (2024). Narratives in Economics. *JEL*.
- Paul, J. & Criado, A. (2020). Art of Writing Literature Reviews. *APJM*.

Notes — Slide 32

Backup references for Q&A. These are the most-cited papers in the review and the methodological anchors for the algorithmic framework.

Backup: Full PRISMA Flow



Notes — Slide 33

Detailed PRISMA flow for methodology questions. Shows the complete chain from 288 initial papers to 16 final included papers, with exclusion reasons at each stage. Compliant with PRISMA 2020 reporting standards.

Backup: Stability Check Details

PCA Threshold Sensitivity

Threshold	Components	Papers	Avg Rel.
80%	2	26	0.507
85%	3	26	0.507
90%	3	26	0.507
95%	4	26	0.507
98%	4	26	0.507
99%	4	26	0.507

Complete stability: identical paper set across all thresholds.

Algorithm Comparison (Full)

Method	Papers	Avg Rel.	Silhouette
K-Means	26	0.504	0.347
GMM	<26	>0.504	<0.347
Agglom. (Ward)	50	—	>0.347

K-Means selected for:

- Best balance of precision and recall
- Reasonable cluster separation
- Not too aggressive (GMM) or permissive (AC)

Notes — Slide 34

Full detail tables for robustness questions. The PCA threshold table is the strongest evidence: regardless of whether you retain 80% or 99% of variance, the exact same 26 papers are selected. The algorithm comparison shows K-Means occupying the sweet spot between GMM (too few papers, potentially missing relevant ones) and Agglomerative Clustering (too many papers, diluting the corpus).

Backup: All 16 Papers

Author(s)	Year	Stream	Method	Key Finding
Tuckett et al.	2014	Modeling	Directed sentiment	Conviction narratives detect FX regime changes
Hu et al.	2021	Understanding	Annotation	Structured labeling for economic opinion/emotion
Hsu et al.	2021	Modeling	Freq. + regression	Historical narratives drove 1930s capital flows
Chen et al.	2022	Modeling	Topic model	COVID narrative virality → volatility transmission
Zmandar et al.	2022	Understanding	Transformer	CoFiF Plus: French financial narrative corpus
Borup et al.	2023	Modeling	Topic model	Narrative expectations outperform sentiment
Sy et al.	2023	Understanding	BERT ensemble	Argument mining in earnings calls
Miori & Petrov	2023	Modeling	GPT + graphs	Narrative fragmentation ≈ systemic risk
Stander	2024	Modeling	FinBERT	SA news sentiment for IFRS 9 credit risk
Agarwal et al.	2024	Modeling	Emotion metrics	Fear/anger predict returns in bub-

Notes — Slide 35

Complete reference table for all 16 papers in the final corpus. Organized by year. Shows the stream (understanding vs. modeling), primary method, and key finding for each paper. Useful for Q&A when audience asks about specific papers or methods.