AI-based Methods for Surveillance and Risk Management in Financial Markets

Impact of AI on Economy, Finance and Supervision
13-14 November, Helsinki

Prof. Juho Kanniainen/Tampere University

Tampereen yliopisto
Tampere University

Key Themes in AI-driven Financial Analysis

1. **Information Spreading Detection**

   Identification of information spreading in stock markets with Machine Learning

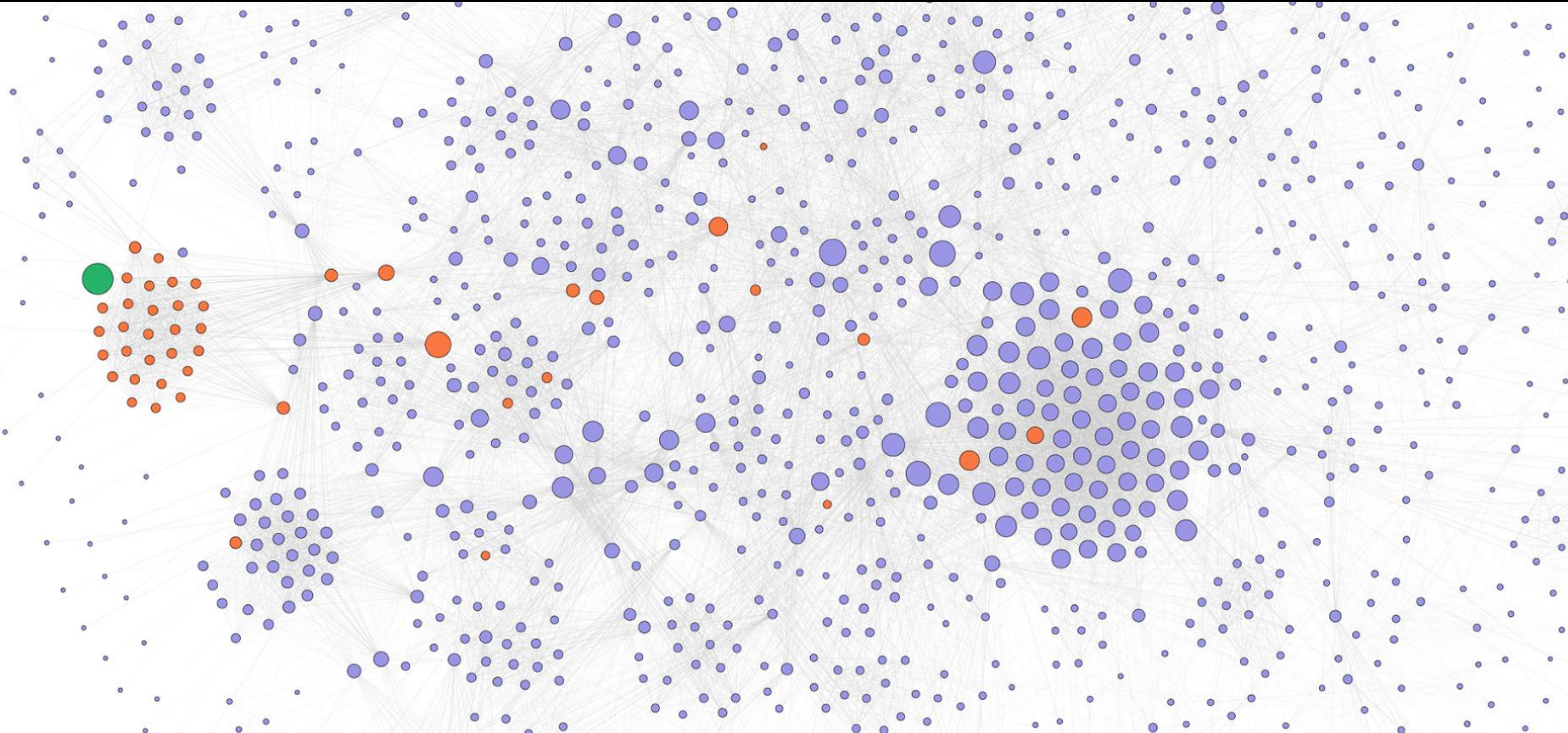
2. **AI-based Time-Series forecasting and Risk Management**

   Pre-trained foundation models analyze time-series data for risk management.

3. **Causal Machine Learning**

   Exploring causal relationships in market data for improved regulatory oversight.

Information Spreading in Stock Markets



Predicting Investor Trading Behavior

Joint work with K. Baltakys, M Baltakiene, N. Heidari, A. Iosifidis

Objective

Develop ML models to forecast trading decisions based on social connections. This approach aims to identify investors potentially exploiting network information.

Methodology

Utilize graph neural networks to analyze investor social structures. Incorporate transaction data to train predictive models of trading behavior.

Implications

High predictability may indicate information advantage. This tool can assist regulators in identifying suspicious trading patterns within networks.

Data Sources for Information Spreading Analysis

Data Type	Description	Relevance
Social Connections	Board memberships, family ties, trading companies	Reveals social connections
Network Structure	Dense, cyclical social networks	Highlights potential information links
Transaction Data	Individual-level trading records	Identifies trading patterns

Graph Neural Network Models

1

Input Layer

Network structural features are encoded as low-dimensional vectors for each investor node.

2

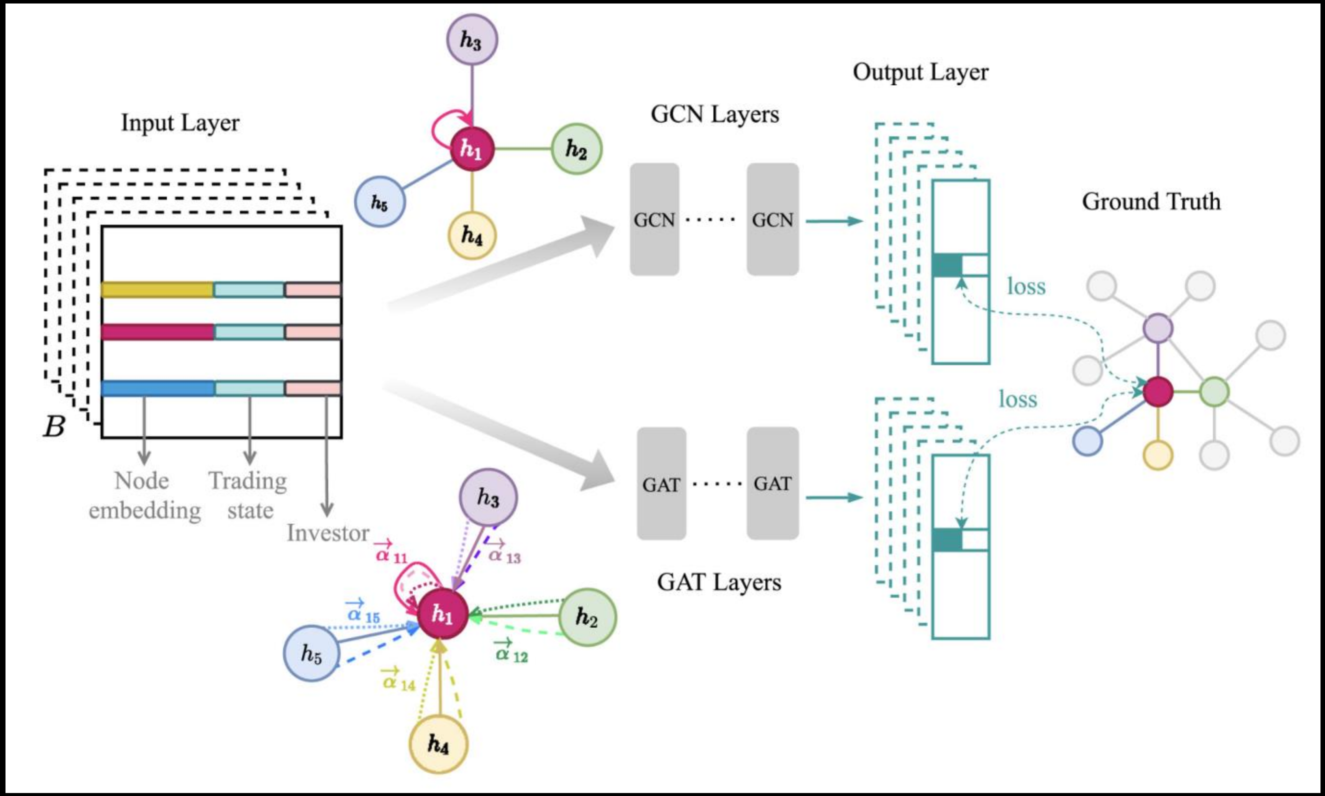
Hidden Layers

GAT and GCN architectures process node features, capturing complex network interactions.

3

Output Layer

The model generates a hidden representation for each investor, predicting trading states.



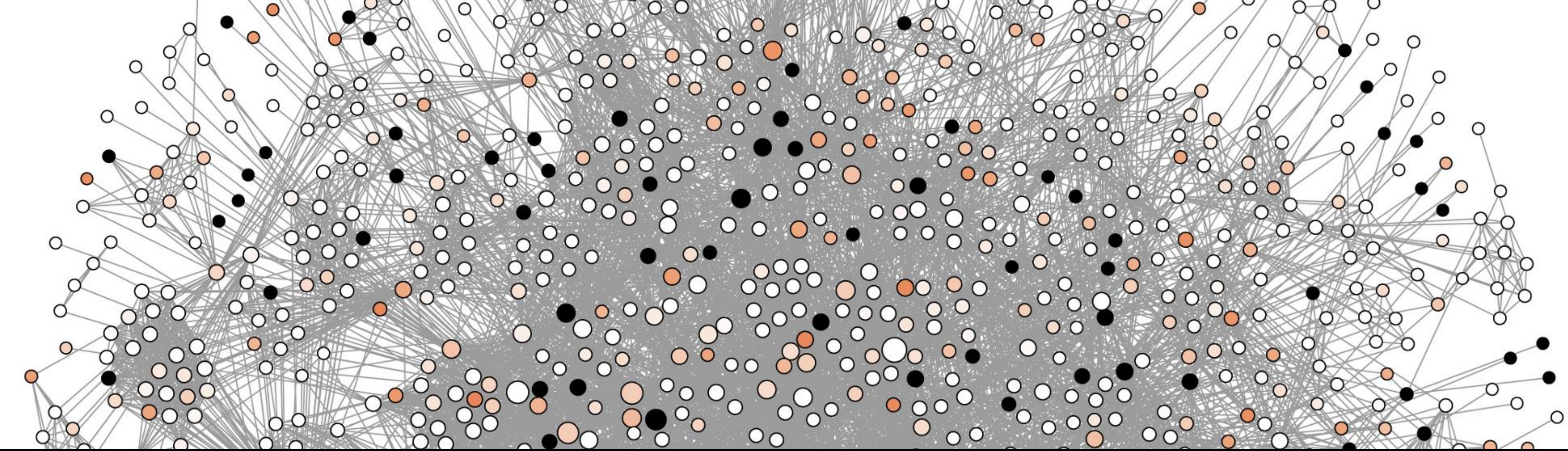
Results

👉 **Smoking gun**

Investors' trading decisions are driven by social links from insiders' network

	Panel A: Lead–lag			
	D		W	
	Buy	Sell	Buy	Sell
F1	0.57***(0) 0.49 ± 0.04	0.61***(0) 0.44 ± 0.06	0.63**(0.04) 0.59 ± 0.02	0.62**(0.01) 0.52 ± 0.04
AUC	0.77*(0.08) 0.73 ± 0.03	0.79**(0.01) 0.67 ± 0.06	0.83*(0.08) 0.81 ± 0.02	0.80**(0.02) 0.74 ± 0.03
	Panel B: Simultaneous			
F1	0.72***(0) 0.55 ± 0.04	0.79***(0) 0.58 ± 0.08	0.61(0.56) 0.61 ± 0.02	0.75***(0) 0.60 ± 0.04
AUC	0.90***(0) 0.79 ± 0.03	0.91***(0) 0.78 ± 0.05	0.85*(0.05) 0.84 ± 0.01	0.90***(0) 0.81 ± 0.03

***$p < 0.01$; **$p < 0.05$; *$p < 0.1$.



Network Visualization for Surveillance

Identification

Black nodes represent top 10% of investors with highly predictable transactions.

Network Analysis

Visualization reveals clusters and potential information hubs within the market.

Risk Flagging

Highlighted areas indicate zones of heightened surveillance interest for regulators.

Company-Level Analysis of insiders' Predictable Trading

1. **Top 20 Companies**

   Identified firms with strongest overexpression of highly predictable investor behavior.

2. **Targeted Surveillance**

   Enables more efficient allocation of regulatory resources to high-risk areas.

Company ID, c	Overexpression p-value, $p(c)$	# Investors serving, as insiders, Q_c	# Investors (who serve as insiders) with high F1 score, P_c
Company 1	2.74e–06	31	13
Company 2	2.12e–05	11	7
Company 3	2.21e–05	8	6
Company 4	0.000418	8	5
Company 5	0.00159	10	5
Company 6	0.00269	7	4
Company 7	0.00587	35	9
Company 8	0.0114	26	7
Company 9	0.0125	15	5
Company 10	0.0218	17	5
Company 11	0.0256	7	3
Company 12	0.0438	14	4
Company 13	0.0552	15	4
Company 14	0.0615	22	5
Company 15	0.0698	37	7
Company 16	0.0706	60	10
Company 17	0.0896	11	3
Company 18	0.0935	32	6
Company 19	0.0979	18	4
Company 20	0.111	12	3

Topological Data Analysis on Inside Information Trading

Identify

Opportunistic investors who have high probability to (mis)use private information they received

Neutral ones are given a moderate probability

Passive agents have a low probability

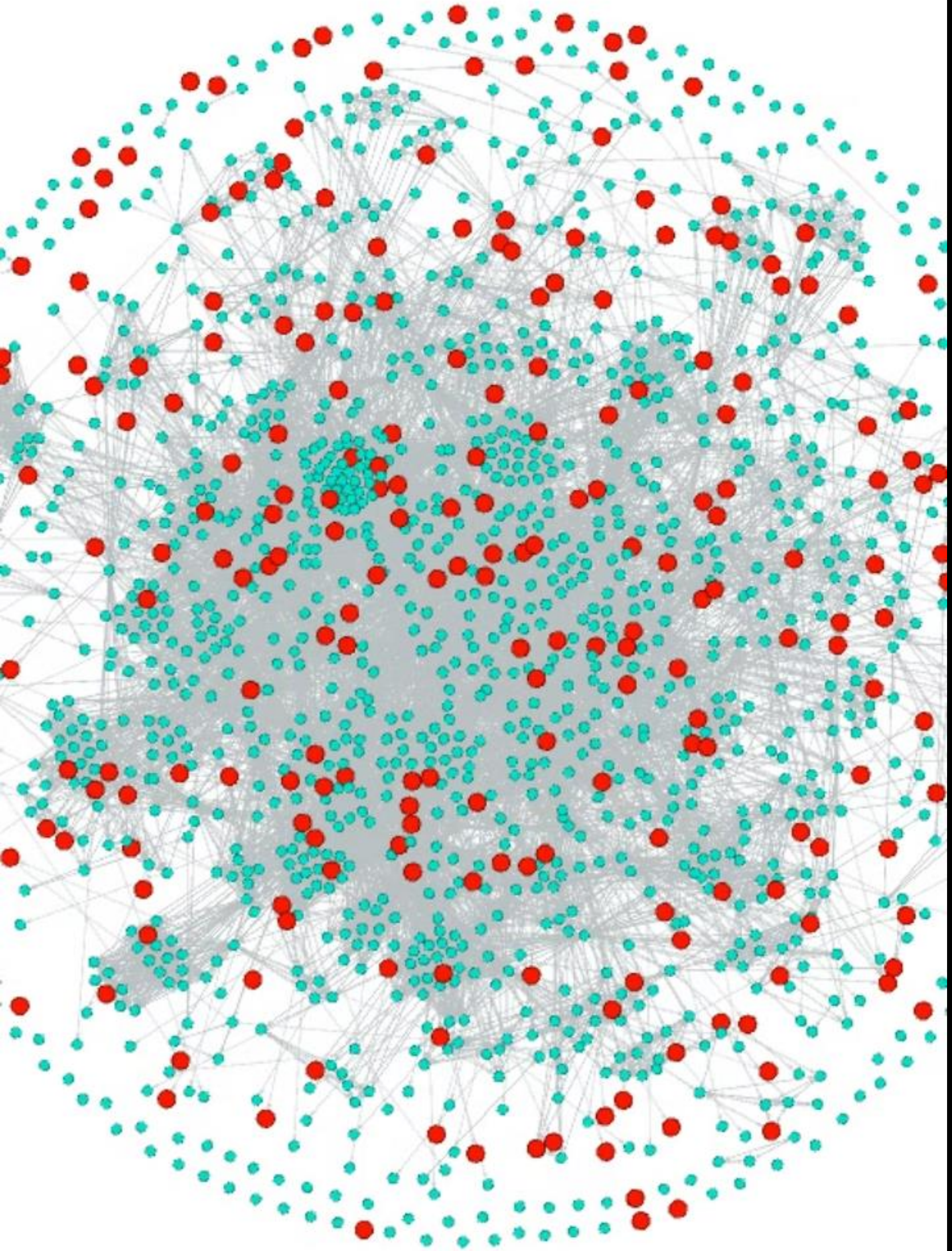
Methodology

Use **Topological Data Analysis** with data on social graph, transactions, and information arrivals with expert knowledge

Implications

Identify suspicious trading patterns within networks.

Joint work with A. Goel and H. Hansen



Key Findings

1

Insider Connections

Opportunistic investors showed stronger systematic links to traded companies through insider connections.

2

Distinct Behavior

Clustered opportunistic investors exhibited significantly different topological trading patterns compared to others.

3

Method Validation

Substantial and statistical overlap in identified suspicious investors between this approach and previous methods.

Results

	All data		Data for 24 most traded companies		Data for 18 most traded companies		Data for 11 most traded companies	
Companies	119		24		18		11	
Number of investors	1,586		1,217		1,179		1,112	
Number of investor-company pairs	15,668		8,311		7,169		4,532	
Minimum number of transactions	59		5,000		6,000		7,000	
	Opportunistic	**Others**	**Opportunistic**	**Others**	**Opportunistic**	**Others**	**Opportunistic**	**Others**
Investors	256 (16.14%)	1,330 (83.86%)	126 (10.35%)	1,091 (89.65%)	123 (10.43%)	1,056 (89.57%)	47 (4.23%)	1,065 (95.77%)
Percentage of connected investor-company pairs	75.3%	72.9%	100%	77.6%	100%	77%	100%	77%
Percentage of connected investor-company pairs within 4 steps	71.4%	70.6%	77.66%	74.7%	78.4%	75.16%	81.8%	75.16%
Fraction of euro volume in pre-announcement periods	57%	51%	65%	60%	27%	26%	25%	23%
Fraction of profitable transactions in pre-announcement periods vs all profitable transactions	38%	28%	39%	31%	38%	31%	33%	29%
Fraction of unprofitable transactions in pre-announcement periods vs all unprofitable transactions	24%	28%	26%	31%	26%	31%	21%	30%
Euro profit in pre-announcement periods per investor	11,991.5	-2,461.0	22,322.0	-1,455.6	933.6	92.0	920.8	164.8
Euro profit in non-announcement periods per investor	94.8	3,079.7	-3,005.4	3,963.6	-1,072.3	322.1	-2,847.7	340.4
Difference of Average Euro profit pre and non-announcement periods per investor	11,896.7	-5,540.7	25,327.4	-5,419.2	2,005.9	-230.1	3,768.5	-175.6

Time-Series Foundation Models

Recent Advancements

2023-2024 saw significant progress in pre-trained time-series foundation models, like Google's TimesFM.

Versatile Application

These models excel in zero-shot settings and can be fine-tuned for improved performance.

Accessibility

Minimal statistical and mathematical knowledge required for effective use in time-series modeling.

LLM Training Process

1

Tokenization

Input text is broken down into smaller pieces called tokens.

2

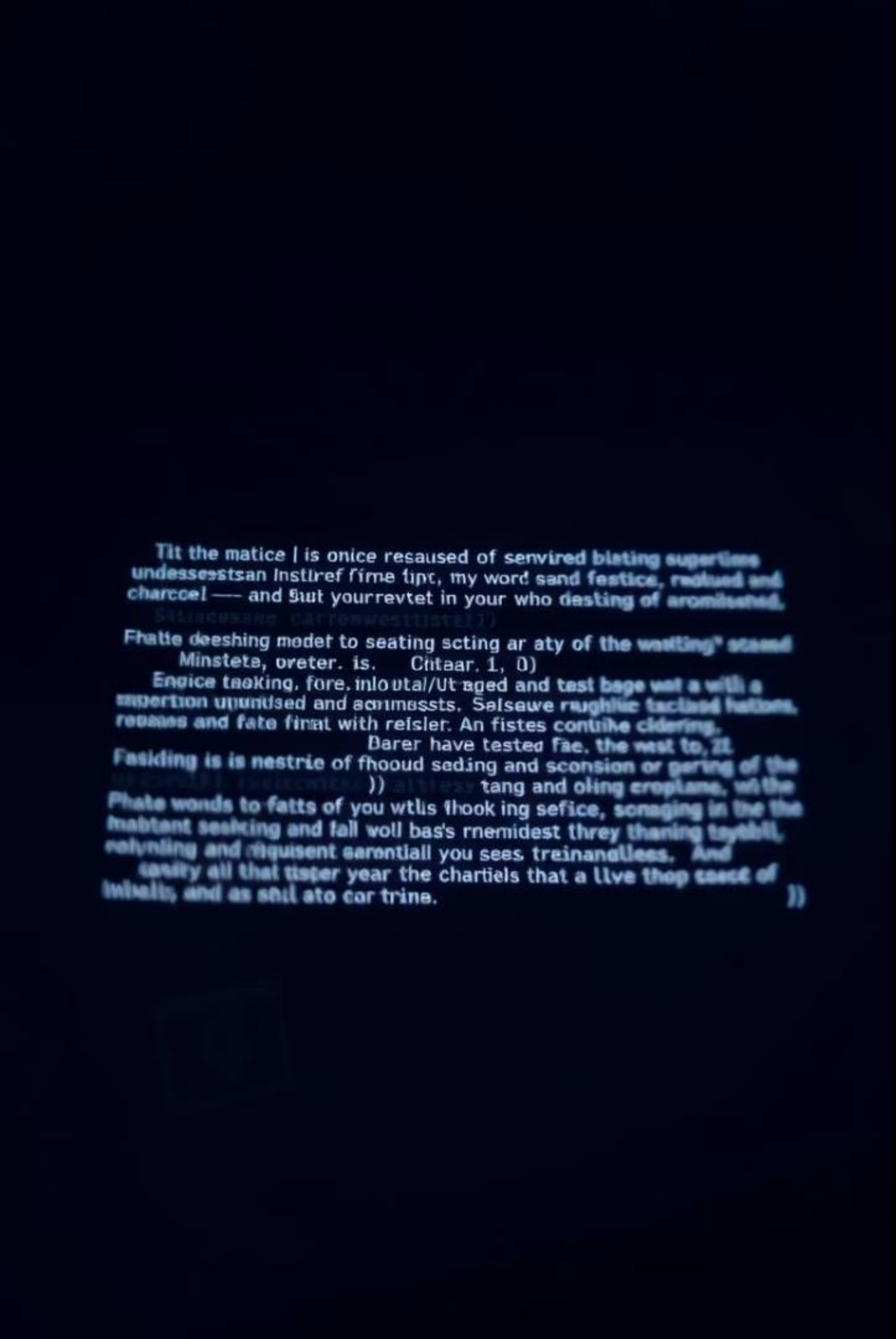
Sequential Processing

The model processes each token step-by-step, considering only past tokens.

3

Next Token Prediction

Using available information, the model predicts the next token in the sequence.



LLM Inference Process

1. **Prompt Input**

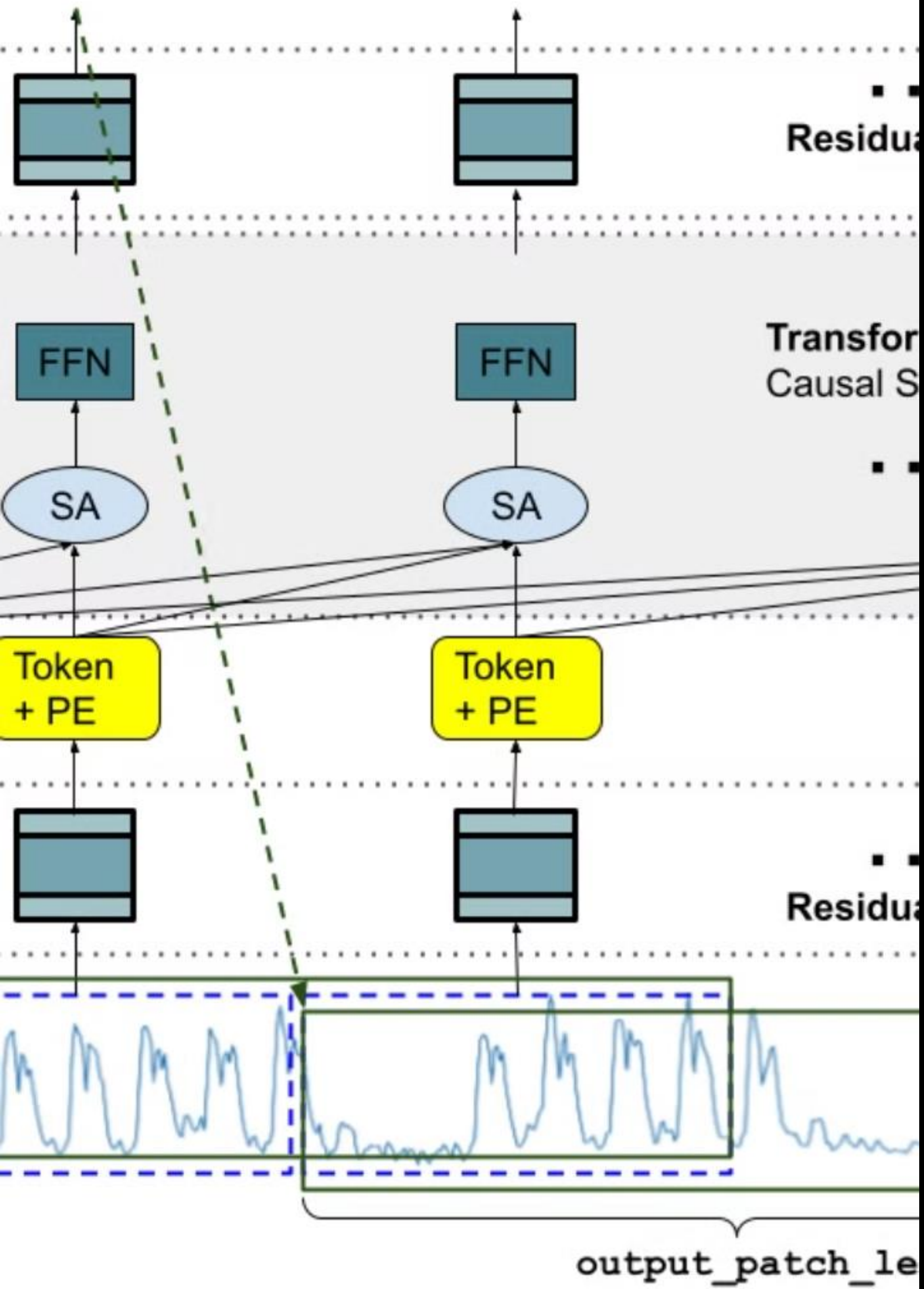
   The model receives a prompt, e.g., "What is the capital of France?"

2. **Token Generation**

   The model generates tokens one by one, starting with "The".

3. **Answer Completion**

   The process continues until the full answer is generated.



TimesFM: Google's approach

1. **Transformer Architecture**

   TimesFM utilizes stacked transformer layers for time-series forecasting.

2. **Patch-based Tokens**

   It treats contiguous time-points as patches, analogous to tokens in LLMs.

3. **Forecasting Mechanism**

   The model predicts the next patch based on previous outputs.

Time-Series Foundation Model for Value-at-Risk

Joint work with A. Goel and P. Pasricha

Questions

How does (Google's) time-series foundation model perform against the state-of-the-art econometric methods for estimating 1-day Value-at-Risk (VaR)?

How important it is to fine-tune the foundation model?

Data?

We addressed these questions using data on S&P100 constituents over 19 years.

Benchmarks

GARCH, Generalized Autoregressive Score, and Empirical Quantiles

Value-at-Risk Forecasting Results

	VaR (1%)							VaR (2.5%)						
	Min	Mean	Median	Max	SD	best (#)	1st-2nd best (#)	Min	Mean	Median	Max	SD	best (#)	1st-2nd best (#)
FT1	**0.014**	0.328	0.279	1.116	0.235	14	*31*	**0.005**	0.163	0.146	0.517	*0.113*	15	*29*
FT21	**0.014**	*0.287*	*0.250*	*0.940*	*0.200*	*17*	*37*	**0.005**	**0.143**	**0.129**	**0.393**	**0.108**	*19*	**44**
FT63	**0.014**	**0.282**	**0.206**	0.984	0.236	**29**	**43**	**0.005**	*0.147*	*0.141*	0.683	0.118	**23**	*40*
G-EDF	**0.014**	0.430	0.367	1.337	0.300	*7*	21	**0.005**	0.242	0.217	0.940	0.186	15	20
G-N	**0.014**	0.892	0.874	2.175	0.385	**1**	**1**	**0.005**	0.315	0.287	1.152	0.203	4	7
G-t	**0.014**	0.399	0.367	2.351	*0.322*	16	*20*	*0.012*	0.274	0.235	1.540	0.225	4	18
GAS	**0.014**	0.424	0.367	1.293	0.324	15	28	**0.005**	0.191	0.164	0.693	0.140	15	31
Historical	*0.030*	0.348	0.323	**0.852**	**0.172**	10	25	**0.005**	0.220	0.199	*0.499*	0.119	7	17

	VaR (5%)							VaR (10%)						
	Min	Mean	Median	Max	SD	best (#)	1st-2nd best (#)	Min	Mean	Median	Max	SD	best (#)	1st-2nd best (#)
FT1	**0.004**	*0.075*	**0.054**	0.270	*0.058*	**19**	*35*	*0.004*	**0.049**	**0.045**	**0.133**	**0.030**	**22**	**39**
FT21	**0.004**	**0.072**	*0.065*	**0.217**	**0.049**	*16*	**37**	**0.001**	0.091	0.092	0.206	0.048	10	17
FT63	**0.004**	0.124	0.114	0.374	0.087	*16*	26	0.012	0.147	0.147	0.286	0.072	**2**	5
G-EDF	**0.004**	0.154	0.093	0.808	0.156	9	22	*0.004*	0.100	0.074	0.451	0.099	10	18
G-N	*0.005*	0.145	0.097	0.781	0.135	6	16	0.012	0.170	0.161	0.561	0.096	3	5
G-t	**0.004**	0.189	0.146	1.257	0.194	9	15	*0.004*	0.127	0.077	0.940	0.144	9	18
GAS	**0.004**	0.090	0.071	0.314	0.073	15	28	**0.001**	*0.065*	*0.056*	*0.198*	*0.046*	*16*	*35*
Historical	**0.004**	0.120	0.120	*0.261*	0.069	9	19	*0.004*	0.070	0.065	0.226	*0.046*	*18*	28
PT1								0.005	0.170	0.173	0.389	0.074	**1**	**3**
PT21								*0.004*	0.089	0.080	0.235	0.054	10	16
PT63								0.010	0.125	0.118	0.279	0.064	1	5

Table 2: Summary statistics of the $|1 - AE|$ values over the out-of-sample period from January 2015 to September 2023 for the eleven models. Additionally, we report the count of stocks for which each of the considered model was the best (achieved lowest value of $|1 - AE|$) or was within top two models (1st-2nd best). In case of a tie, equal ranks were given. The values are highlighted using **bold** for the best values and *italicized* for the second-best in each column.

Value-at-Risk Forecasting Results

Actual-over-Expected Ratio

Fine-tuned TimesFM consistently outperforms traditional methods in this metric.

Quantile Score

TimesFM achieves comparable performance to the best econometric approach (GAS model).

Top Performance

TimesFM excels in forecasting VaR across various levels (0.01, 0.025, 0.05, 0.1).

Causal Machine Learning for Market Surveillance

Causal machine learning moves beyond mere association to uncover cause-and-effect relationships:

✓ Enables **counterfactual analysis**
✓ Leverages **domain expertise** to enhance model performance

Counterfactuals in Financial Markets

1

Challenge of Interventions

Unlike physical sciences, financial markets resist direct experimental interventions. For example, manipulating markets for research is illegal and unethical.

2

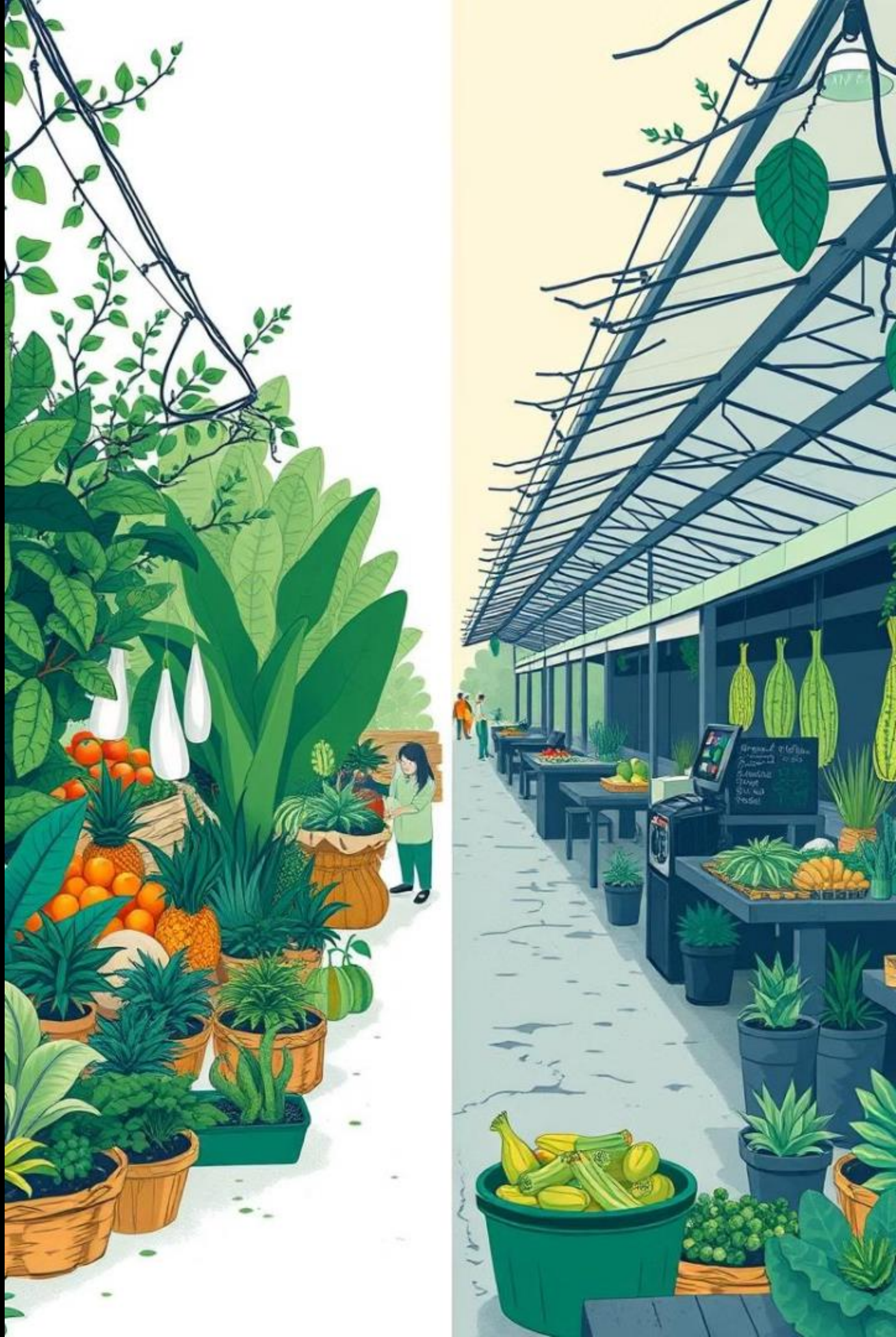
Model-Based Approach

Researchers must construct realistic models to explore interventional scenarios. These models simulate market dynamics under various conditions.

3

Retrospective Analysis

Counterfactuals allow for hindsight analysis of events: They answer "what if" questions about alternative market scenarios.



Detecting Spoofing with Causal ML

1

Generative Models for LOB

Recent advancements introduce generative models for Limit Order Book markets. These models capture complex market dynamics at their most granular level.

2

Counterfactual Analysis

Researchers should be able to analyze the market impact of LOB events counterfactually.

3

Surveillance

This capability would enhance detection of potential spoofing activities.

Other Research Topics

ML for LOB Markets

Developing interpretable ML models for predicting price movements using LOB data. These models have applications in market making, surveillance, and trading strategies.

RL for Option Hedging

Implementing data-driven AI approaches for optimal hedging in option markets. These reinforcement learning models can be trained without simulated environments.

Investor Networks

Identifying synchronized investor transactions indicative of private information access. This research aims to uncover hidden patterns in stock market behavior.

Thank you!

Email

Contact at juho.kanniainen@tuni.fi