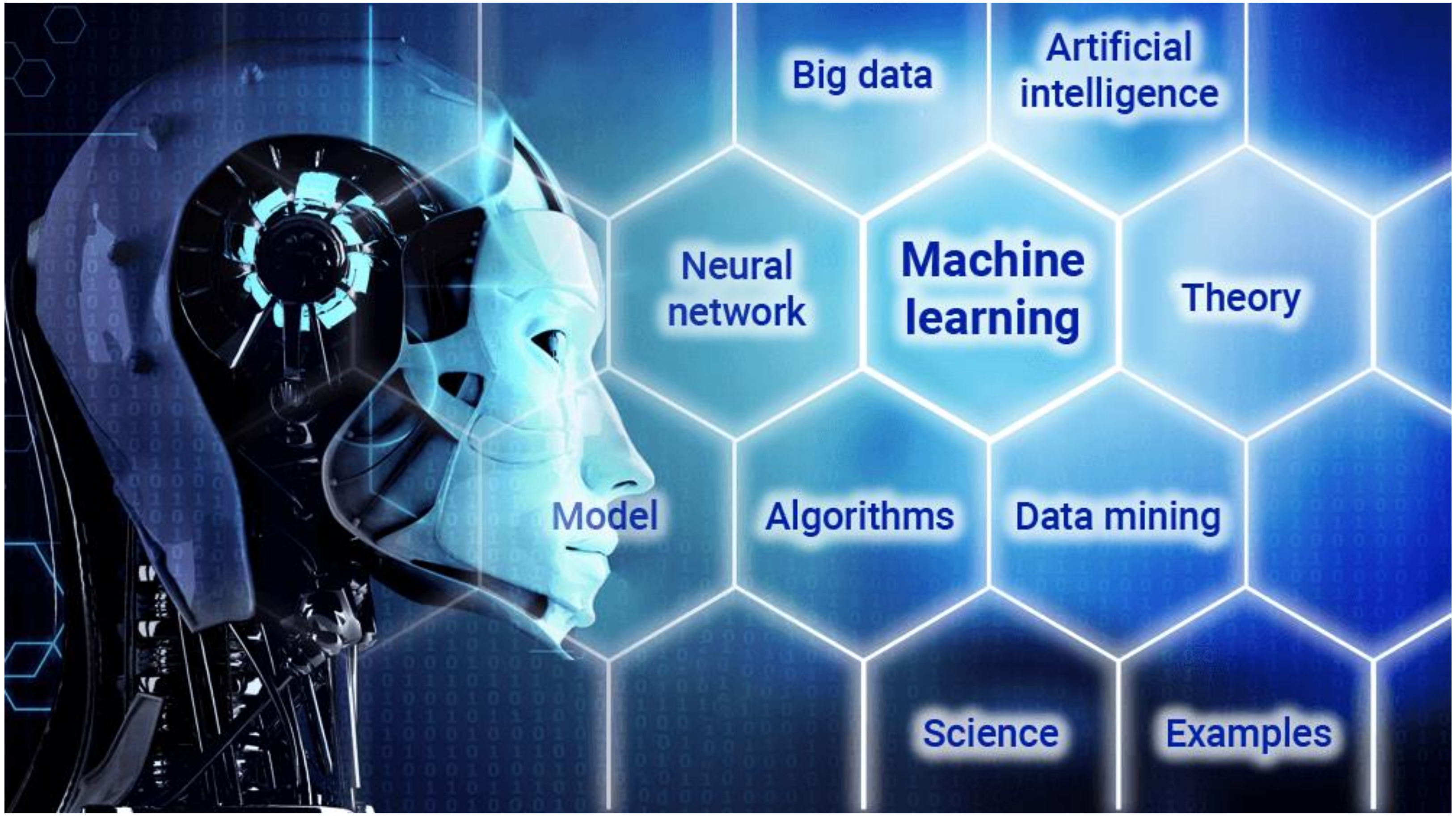
Artificial Intelligence in Equity Investment

Haifeng YOU

2023.11.27

AI, Machine Learning, Big Data, Neural Network…

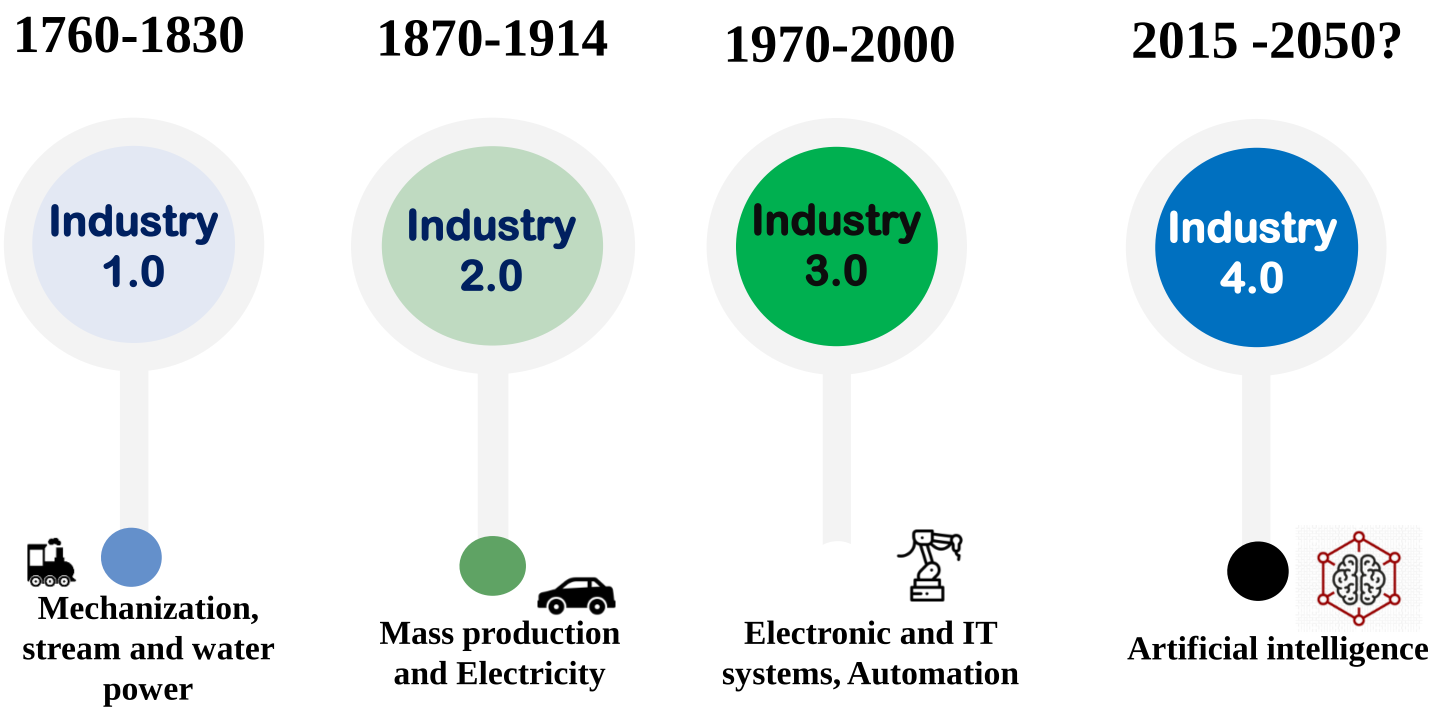


"Artificial intelligence, especially machine learning, is the most important general-purpose technology of our era."

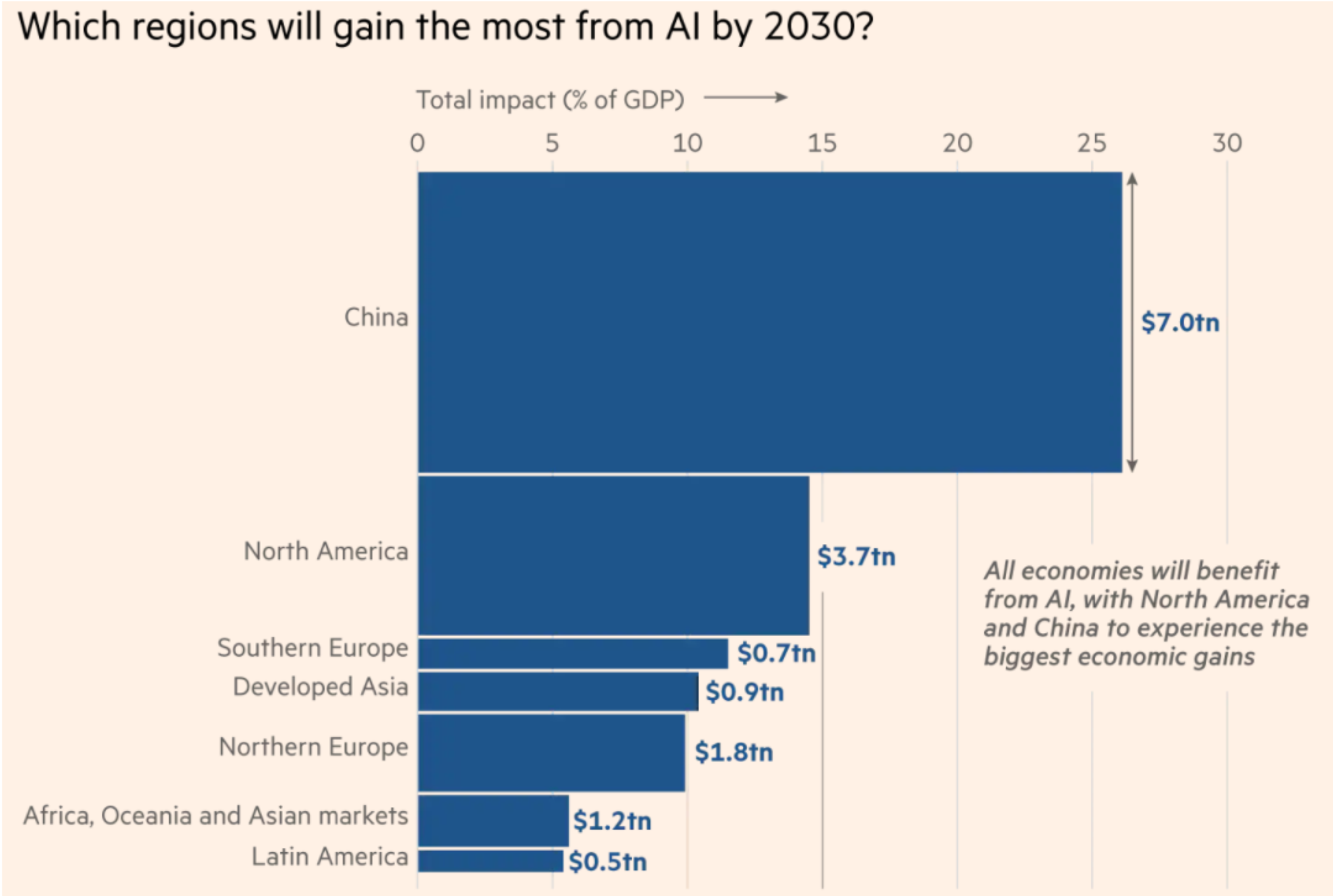
---Erik Brynjolfsson & Andrew McAfee (Harvard Business Review, 2017)

Artificial Intelligence and Industry 4.0

General-purpose technologies (GPTs) are technologies that can affect an entire economy (usually at a national or global level).

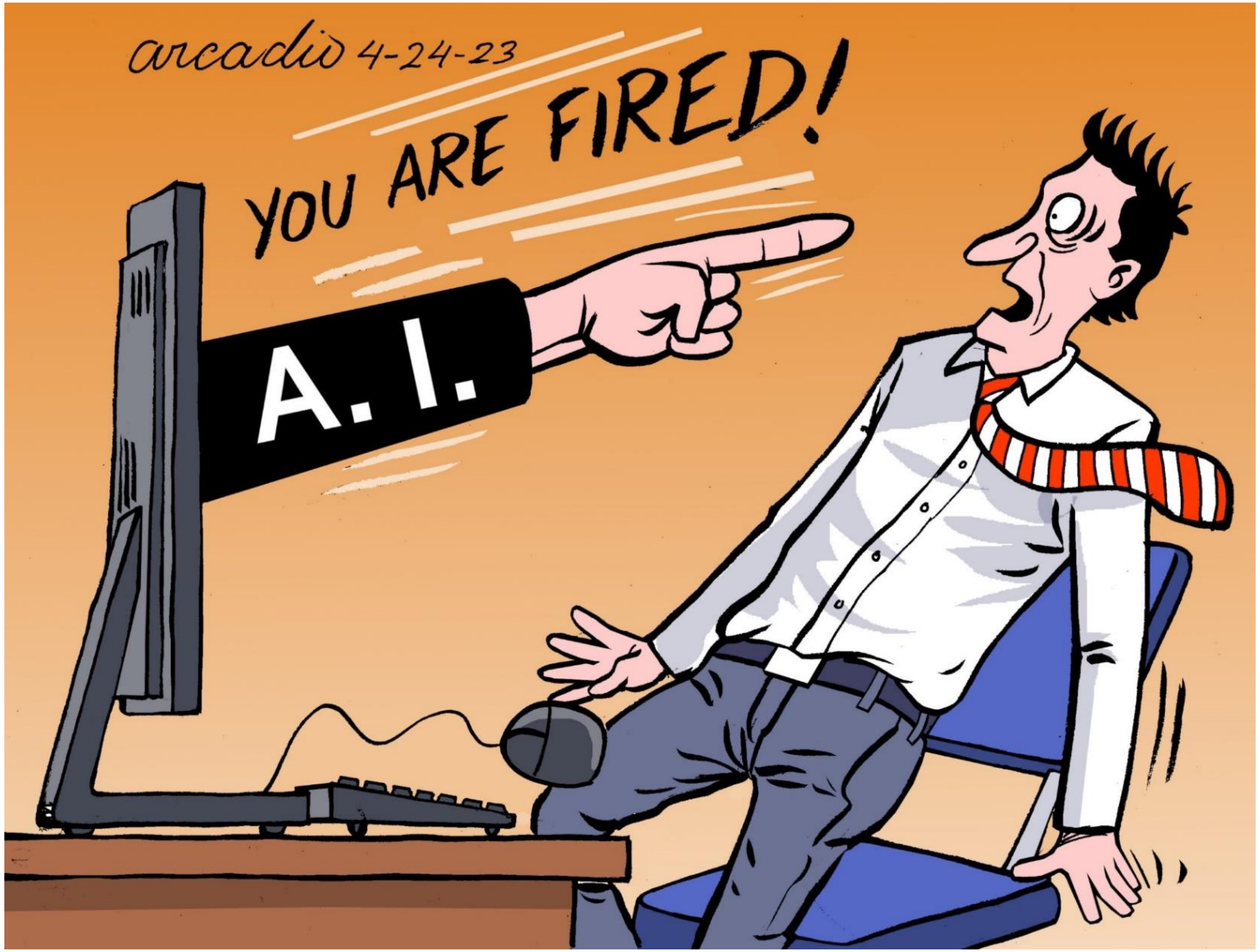


Macroeconomic Impact of AI



The macroeconomic impact of artificial intelligence PWC, 2018

AI Opportunities … (and Risk)



AI Opportunities … (and ????)

Technology

OpenAI researchers warned board of AI breakthrough ahead of CEO ouster, sources say

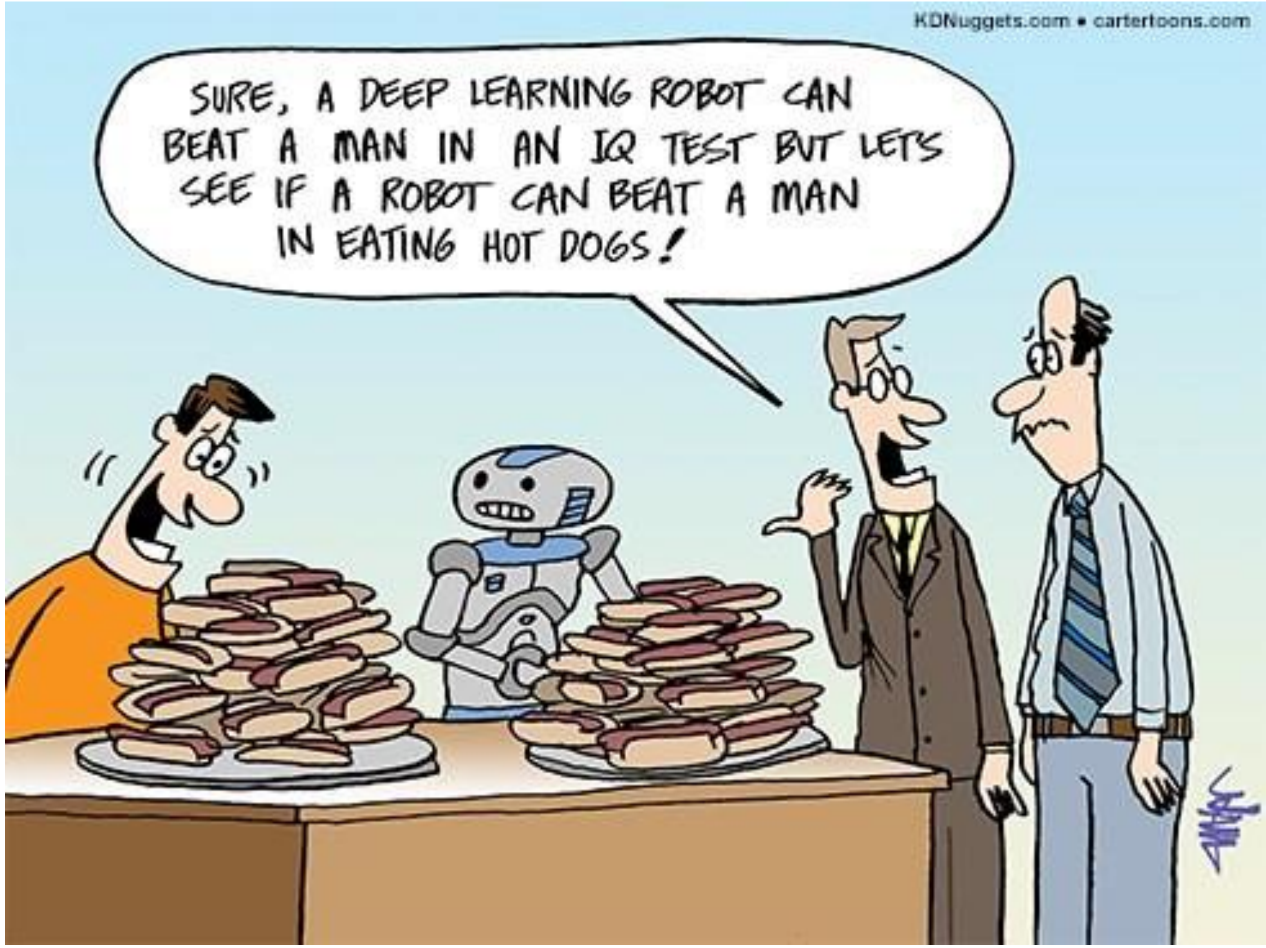
By **Anna Tong**, **Jeffrey Dastin** and **Krystal Hu**

November 23, 2023 5:52 PM GMT+8 · Updated 4 days ago

Aa

Nov 22 (Reuters) - Ahead of OpenAI CEO Sam Altman's four days in exile, **several staff researchers wrote a letter to the board of directors warning of a powerful artificial intelligence discovery that they said could threaten humanity**, two people familiar with the matter told Reuters.

Human Intelligence VS Artificial Intelligence



Human Intelligence VS Artificial Intelligence: Pros

Artificial Intelligence

- Ability to simulate human behavior and cognitive processes
- Capture and preserve human expertise
- Fast response: comprehend large amounts of data quickly.

Human Intelligence

- Intuition, Common sense, Judgement, Creativity, Beliefs etc
- The ability to demonstrate their intelligence by communicating effectively
- Plausible reasoning and critical thinking

Human Intelligence VS Artificial Intelligence: Cons

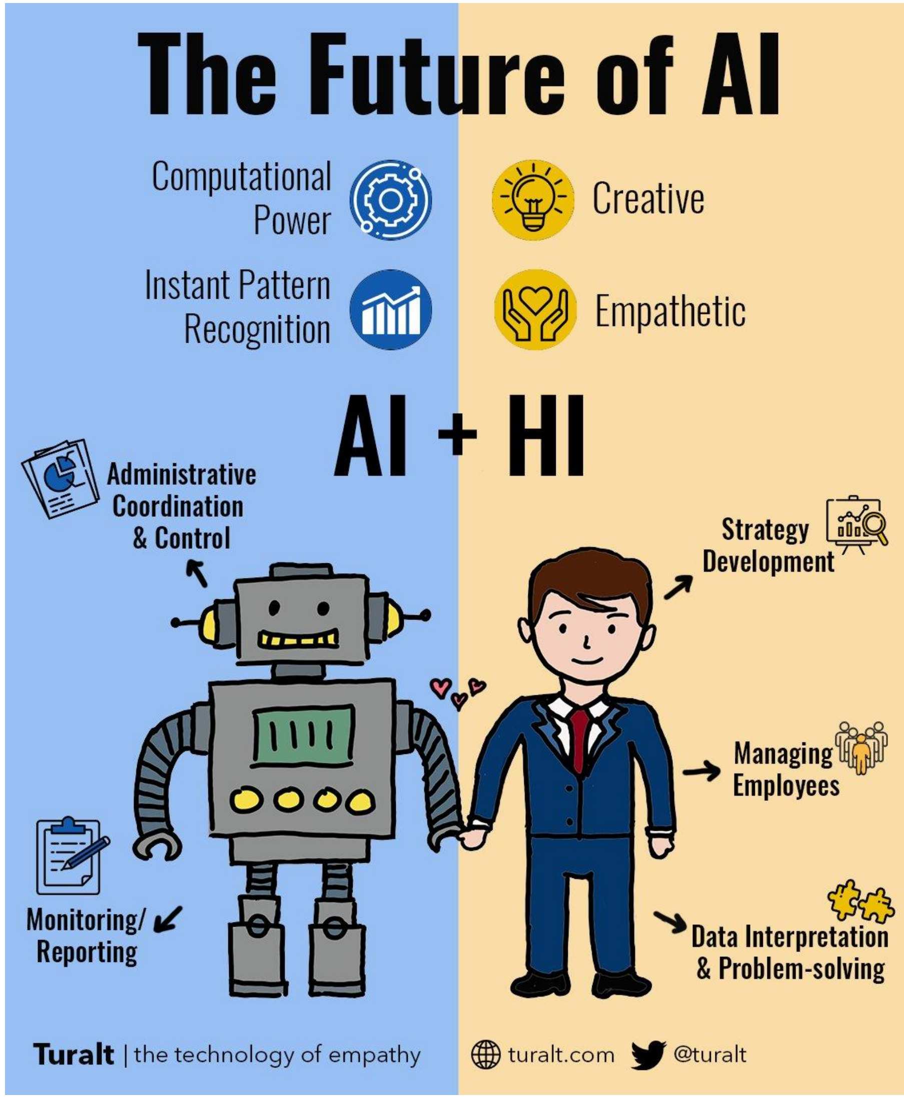
Human Intelligence

- Humans are fallible
- Limited knowledge bases
- Information processing of serial nature proceed very slowly in the brain as compared to computers
- Humans are unable to retain large amounts of data in memory.

Artificial Intelligence

- Lack of creativity, emotion and empathy
- Cannot readily deal with “mixed” knowledge
- May have high development costs
- Raise legal and ethical concerns

The Future of AI (or human society?)



Evolution of Investment Paradigms

Traditional Investing:
Human Intelligence

- Primarily rely on human judgement
- Adaptive to new environment
- Forward looking

- Cognitive constraints
- Emotional swings
- Concentrated portfolios

Quant Investing:
Hardcoded programs

- Primarily rely on expert system
- Efficient information processing
- Immune from human emotions
- Rigorous risk management

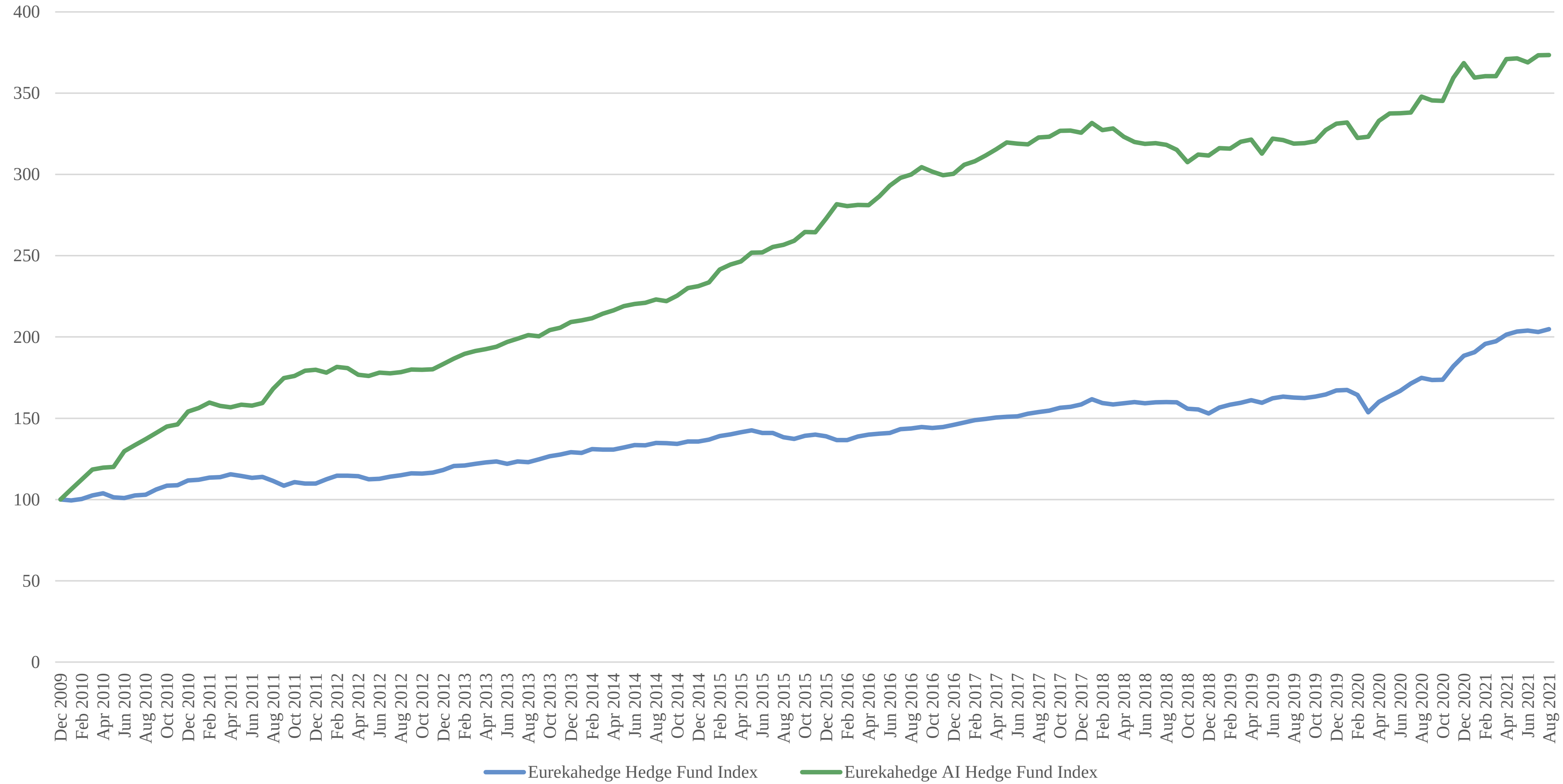
- Static/rigid models
- No adaptability and learning

AI Investing:
AI, Big data & Economics

- Primarily rely on AI/ML
- Utilize both structured and unstructured data
- Adaptability and ability to "learn"

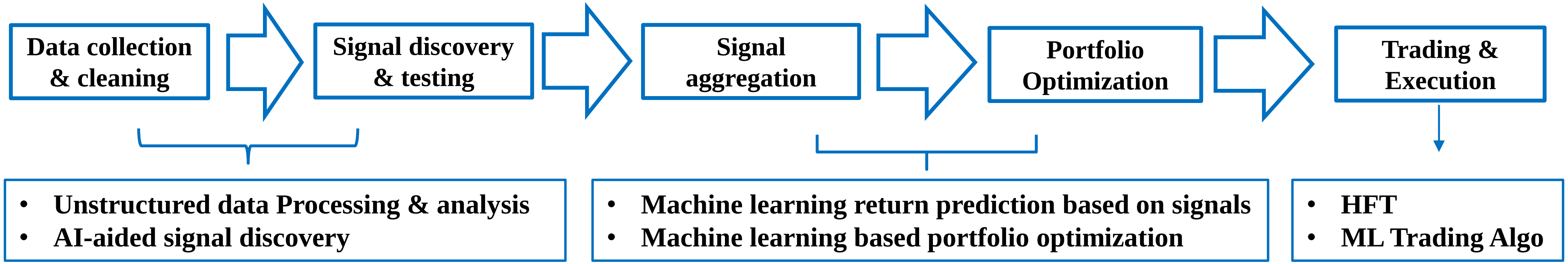
- Overfitting risk
- Low interpretability
- Steep learning curve

AI vs. Traditional Hedge Funds

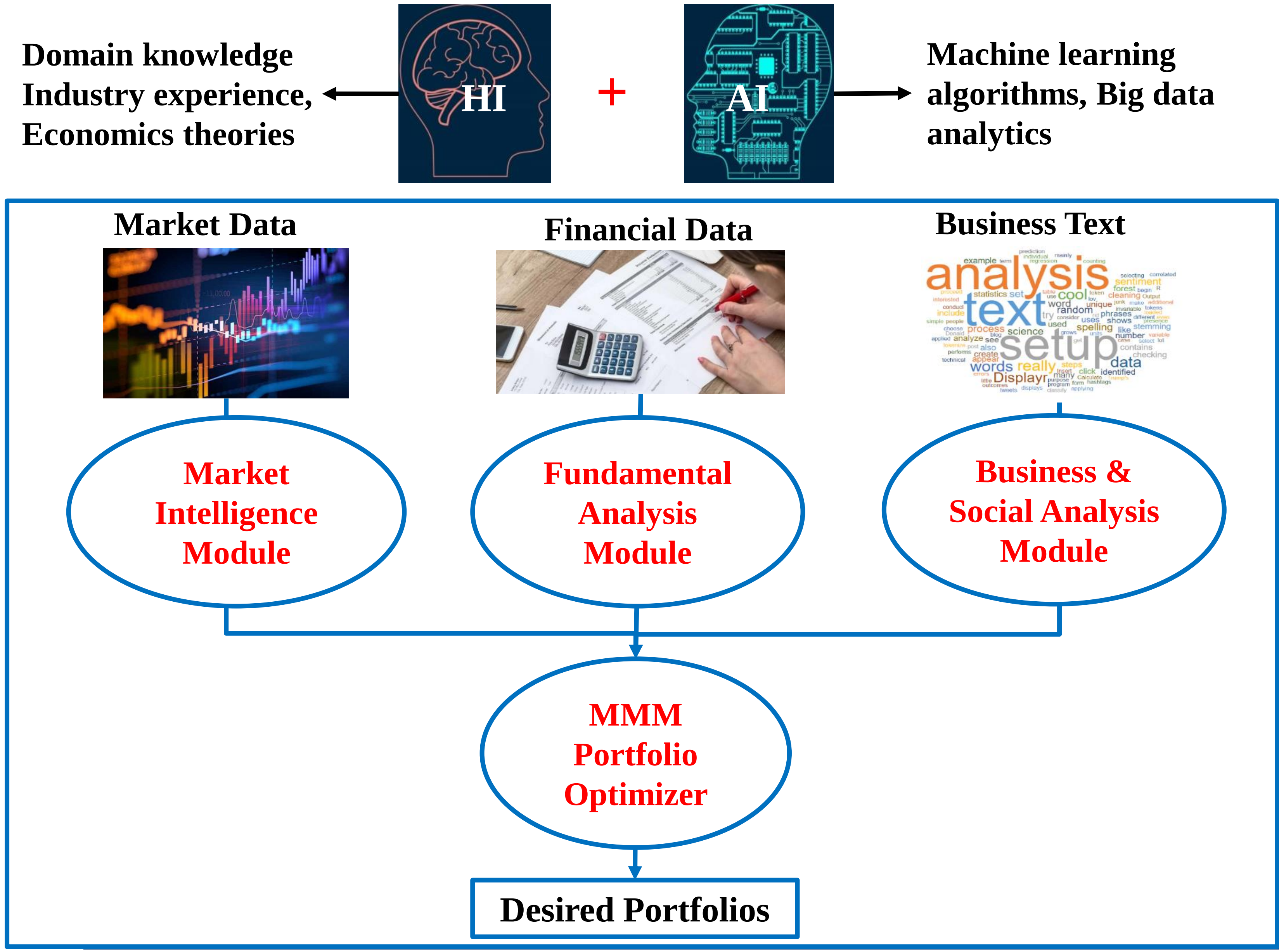


Source: Eurekahedge. The Eurekahedge AI Hedge Fund Index (Bloomberg Ticker - EHFI817) is an equally weighted index of 14 constituent funds. The index is designed to provide a broad measure of the performance of underlying hedge fund managers who utilize artificial intelligence and machine learning theory in their trading processes.

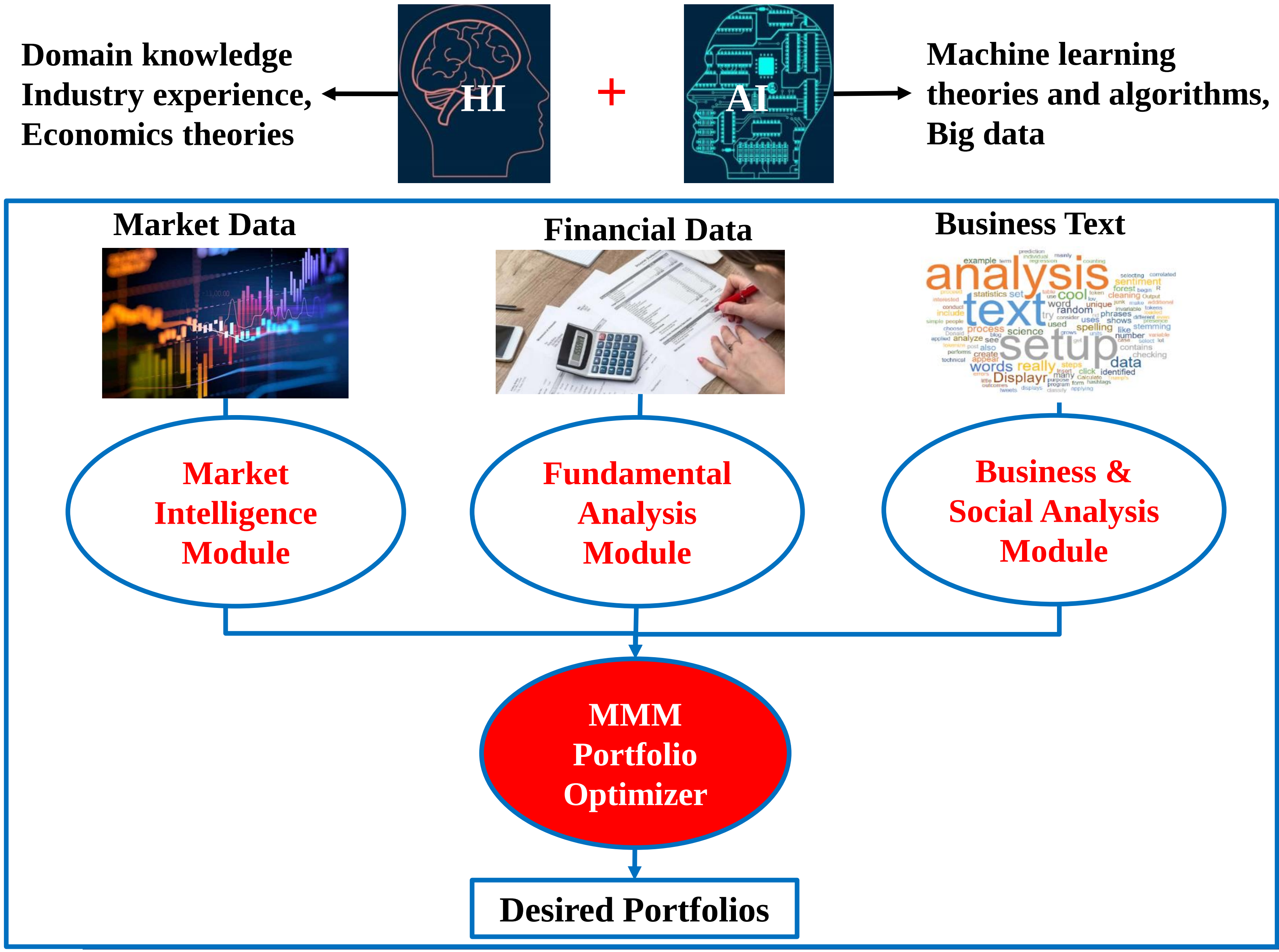
AI and Quantitative Investing



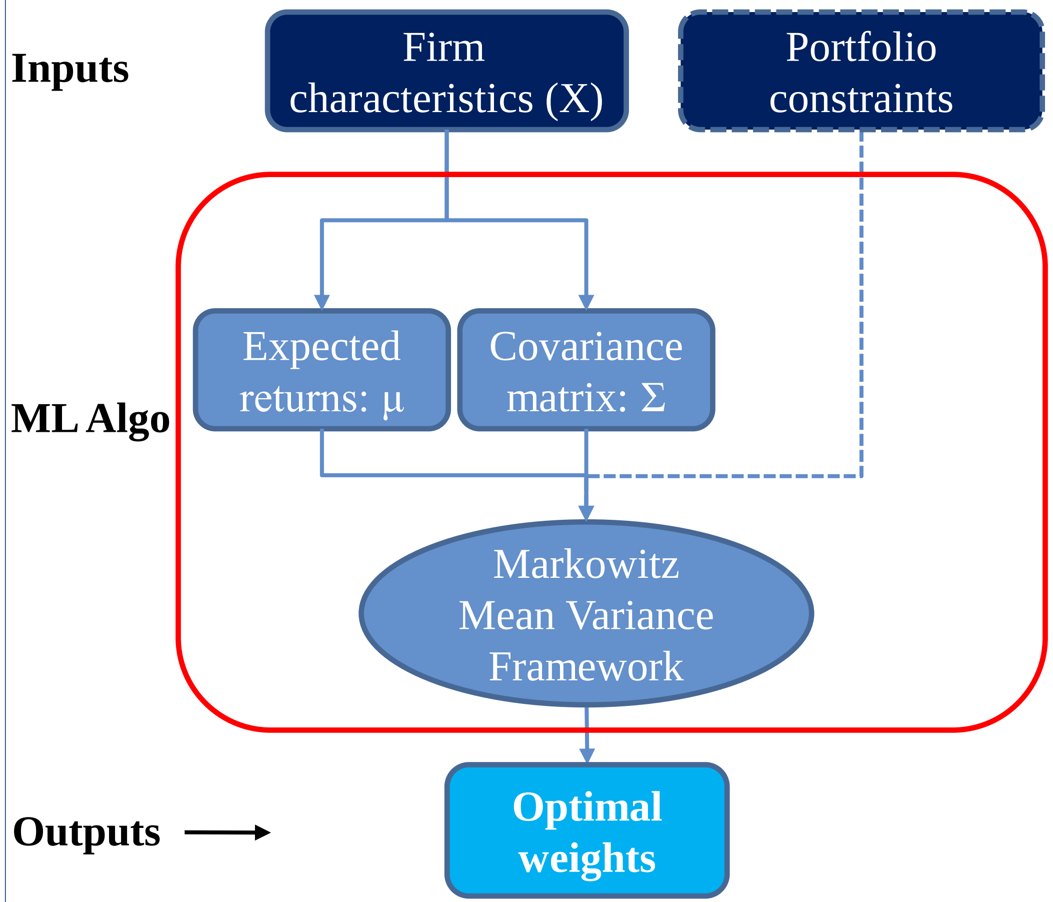
Our AI Investing System



The Core of the AI Investing System: MMM Portfolio Optimizer

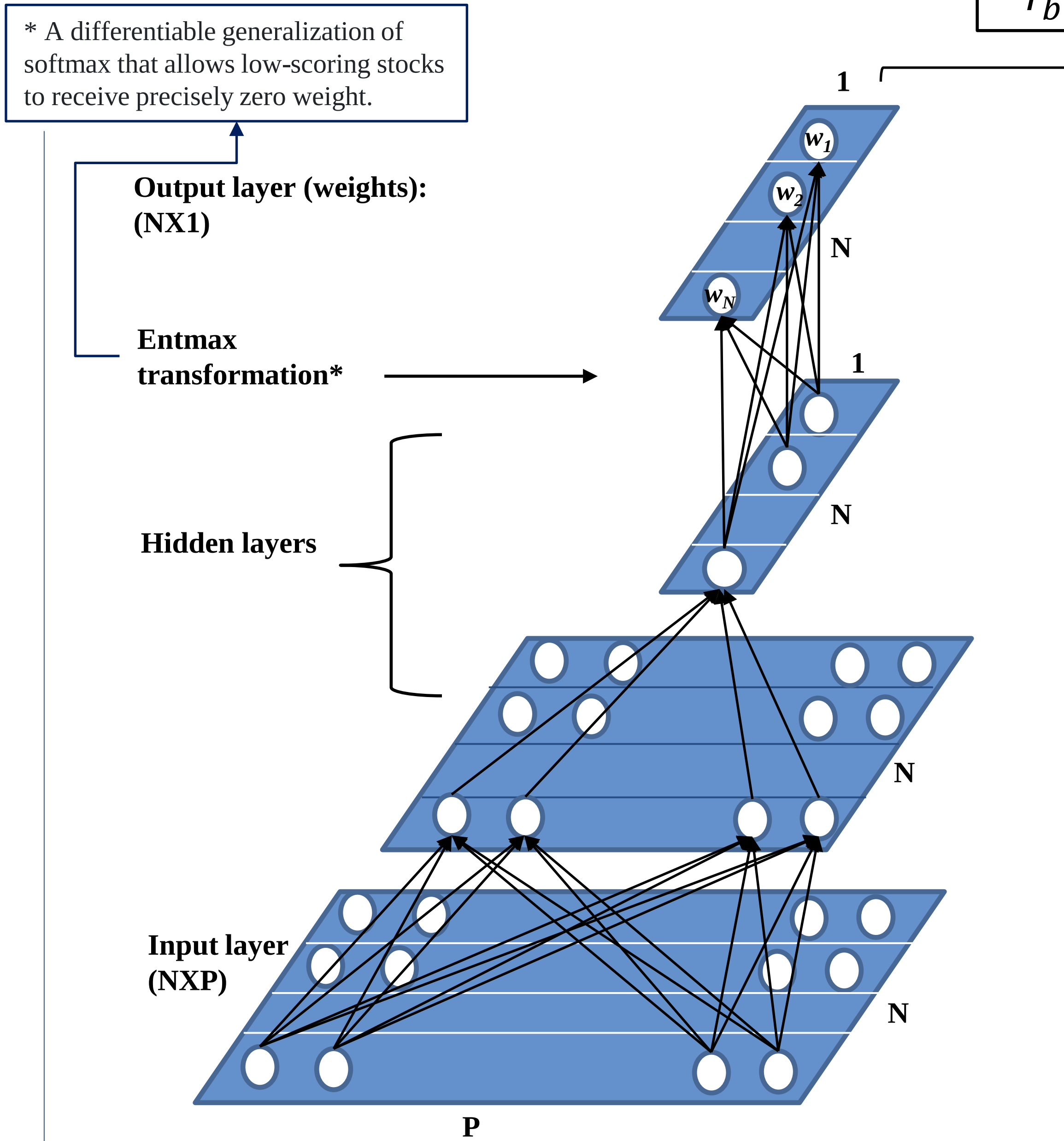


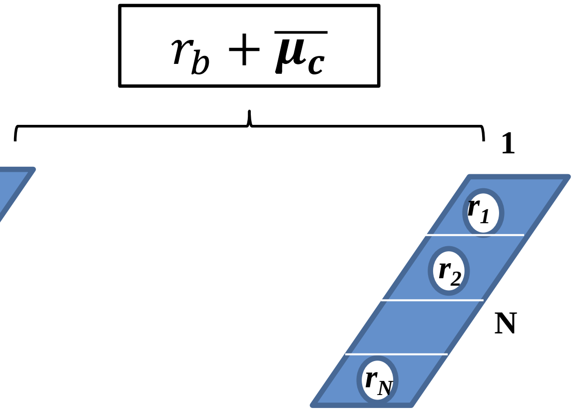
MMM Portfolio Optimizer: When Markowitz Meets Machine



- Theoretically equivalent to Markowitz
- Model optimal portfolio weights as a flexible function of firm characteristics
- Circumvent the needs of estimating explicit return forecasts and risk models
- Allow all factors to contribute to both return & risk
- Accommodate nonlinearity and interactions
- Incorporate machine learning solutions for noise reduction
- Optimal signal weights (and the implicit alpha and risk models) adjust automatically to different investment objectives, universes, and constraints

Architecture of MMM





- Machine learning process minimizes the sample version of $E[(r_b + \overline{\mu_c}) - w(X,\theta)^T r]^2$, i.e.

$$\underset{w(.,\theta)}{arg\,min} \frac{1}{T} \sum_{t=1}^{T} \left[(r_{b,t} + \overline{\mu_c}) - w(X_{t-1},\theta)^T r_t\right]^2$$

- Weight is a nonlinear function of input features/signals
  - Hidden layers + Softmax/Entmax transformation

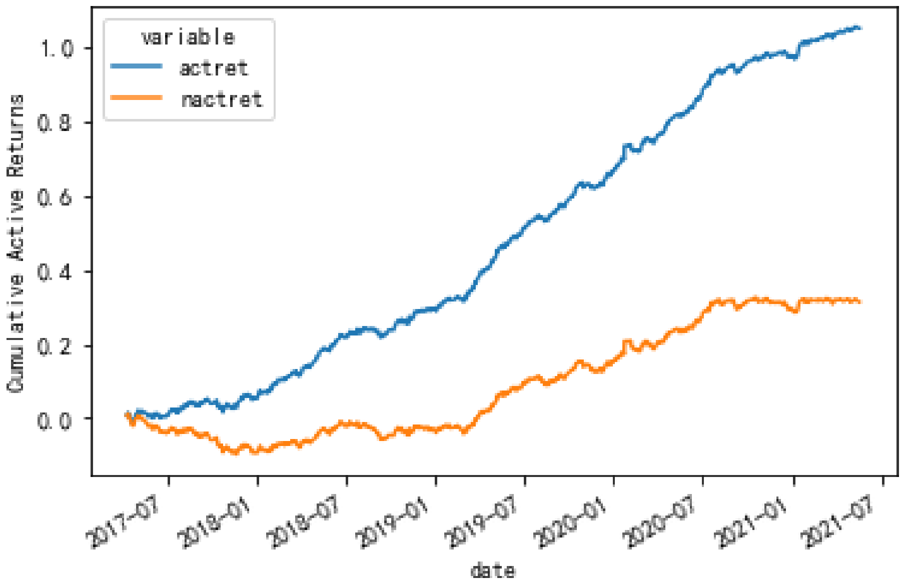
- Entmax ensures long only, full investing and a sparse portfolio: $1^T w = 1, w \geq 0$, n $< N$
  - Constraints are embedded in the network model
  - No need to do the post-optimization truncation and normalization

Traditional Approach with ML vs. MMM Portfolio Optimization (CSI 500 Enhanced Index: 20170411-20210518)

Traditional:

$$\min_{w} \; -\boldsymbol{\mu}_{t+1}^{T}\boldsymbol{w}$$
$$\text{s.t. } \boldsymbol{w} + \boldsymbol{w}^{bmk} \geqslant \mathbf{0}, \mathbf{1}^{T}\mathbf{w} = 0, \boldsymbol{w}^{T}\Sigma\boldsymbol{w} \leq \sigma^{2}$$

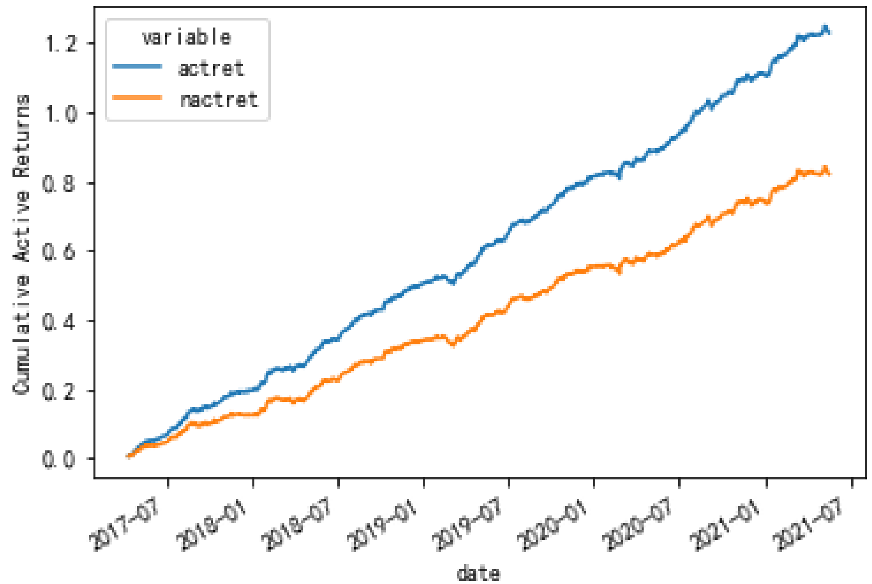
	Annualized Return	**Annualized Volatility**	**Max Drawdown**	**Ratio**
Benchmark	2.81%	23.41%	41.01%	0.12
Total return	29.30%	23.41%	29.64%	1.25
Net return	10.70%	23.39%	42.59%	0.46
Total active return	26.49%	6.22%	3.64%	4.26
Net active return	7.89%	6.19%	10.53%	1.28
Turnover	12,398%			



MMM:

$$\min_{w} \; E[(r_{bmk} + \bar{\mu_c}) - w(X,\theta)^{T} r]^{\mathbf{2}}$$

	Annualized Return	**Annualized Volatility**	**Max Drawdown**	**Ratio**
Benchmark	2.81%	23.41%	41.01%	0.12
Total return	33.80%	23.03%	25.03%	1.47
Net return	23.50%	23.01%	28.22%	1.02
Total active return	30.99%	4.84%	2.63%	6.40
Net active return	20.69%	4.82%	2.84%	4.29
Turnover	6867%			



Data is the New Oil!

Financial Data

- Fundamental prediction
- Financial fraud detection
- Equity valuation

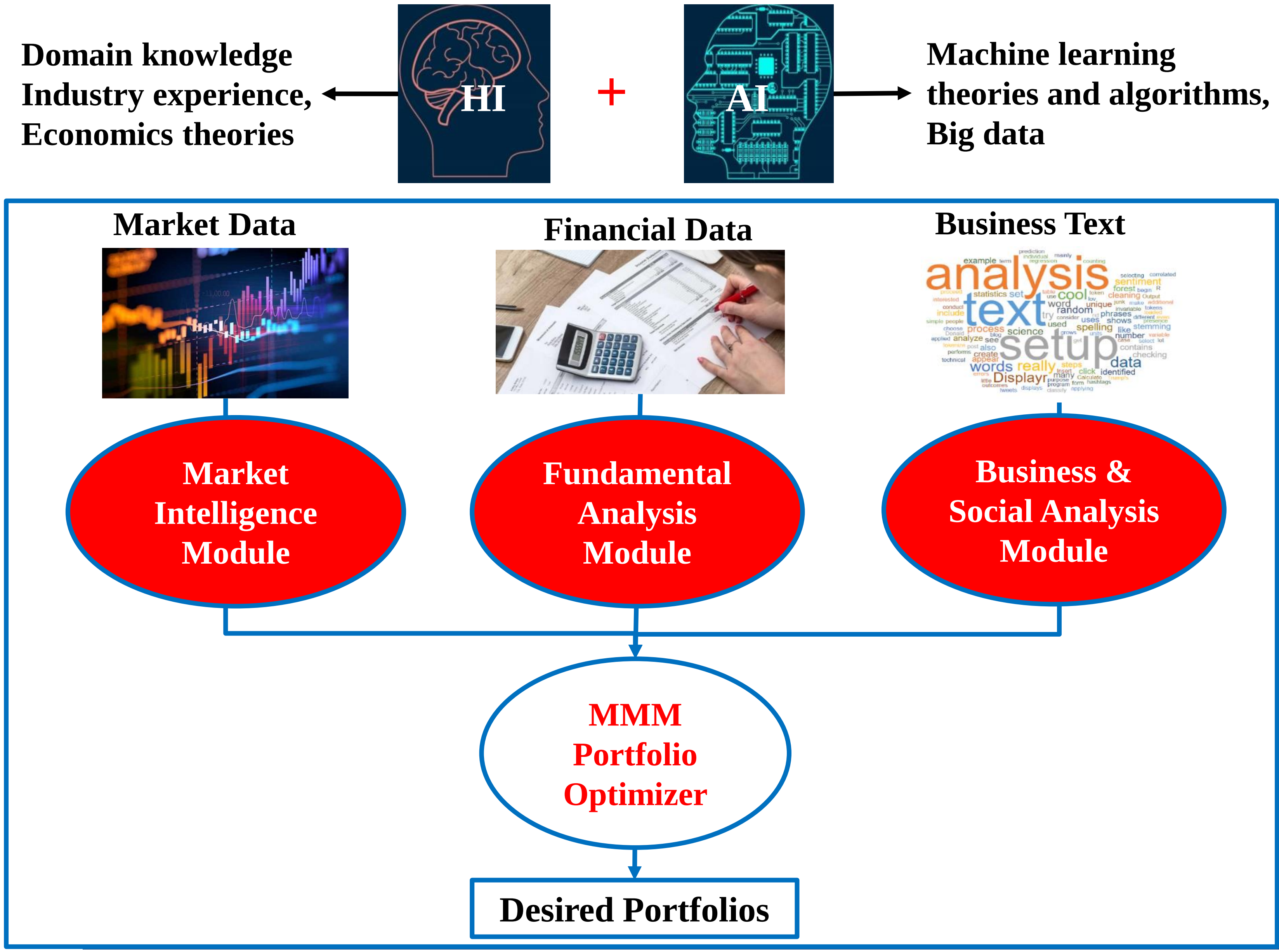
- Technical analysis with ML
- Pattern recognition strategy

Secondary Market Data

- Analyst reports, financial news, blogs, corporate filings
- Sentimental analyses, analyst bias prediction, economic linkage modeling

Business Text

The AI Investing System:
Machine Learning & Big Data Analytics Modules



Outputs of the Machine Learning & Big Data Analytics Modules: Nearly 300 Signals Based on Economic Theories & Machine Learning

Most of the signals can be broadly categorized in the following four groups. Below we highlight some of them:

Valuation
- Machine learning relative valuation incorporating more fundamental information and nonlinearity
- Absolute valuation based on future ML fundamental forecasts
- AI price forecasts based on firm fundamentals, industry competition and macroeconomic indicators etc.

Quality
- Machine based total factor productivity (TFP)
- Predicted (sustainable) profitability with machine learning
- Financial reporting quality based on accounting theory + machine learning

Technical
- Fundamental momentum based on both financial and textual information
- Cross momentum between economically linked firms
- Machine based trend and pattern recognition signals

Sentiment
- Sentiment of financial news, corporate filings and social media
- Analyst over-optimism in target price and fundamental forecasts based on machine learning models
- Investor recognition/sentiment extracted from mutual fund holding data with machine learning

Fundamental Analysis via Machine Learning
Cao and You (2021)

- Reading financial statements is not an extremely pleasant task for most people

- Is machine learning useful for processing financial statement information and generating better earnings forecasts?

- Are ML earnings forecasts useful for making investment decisions?

Data Collection and Feature Selection

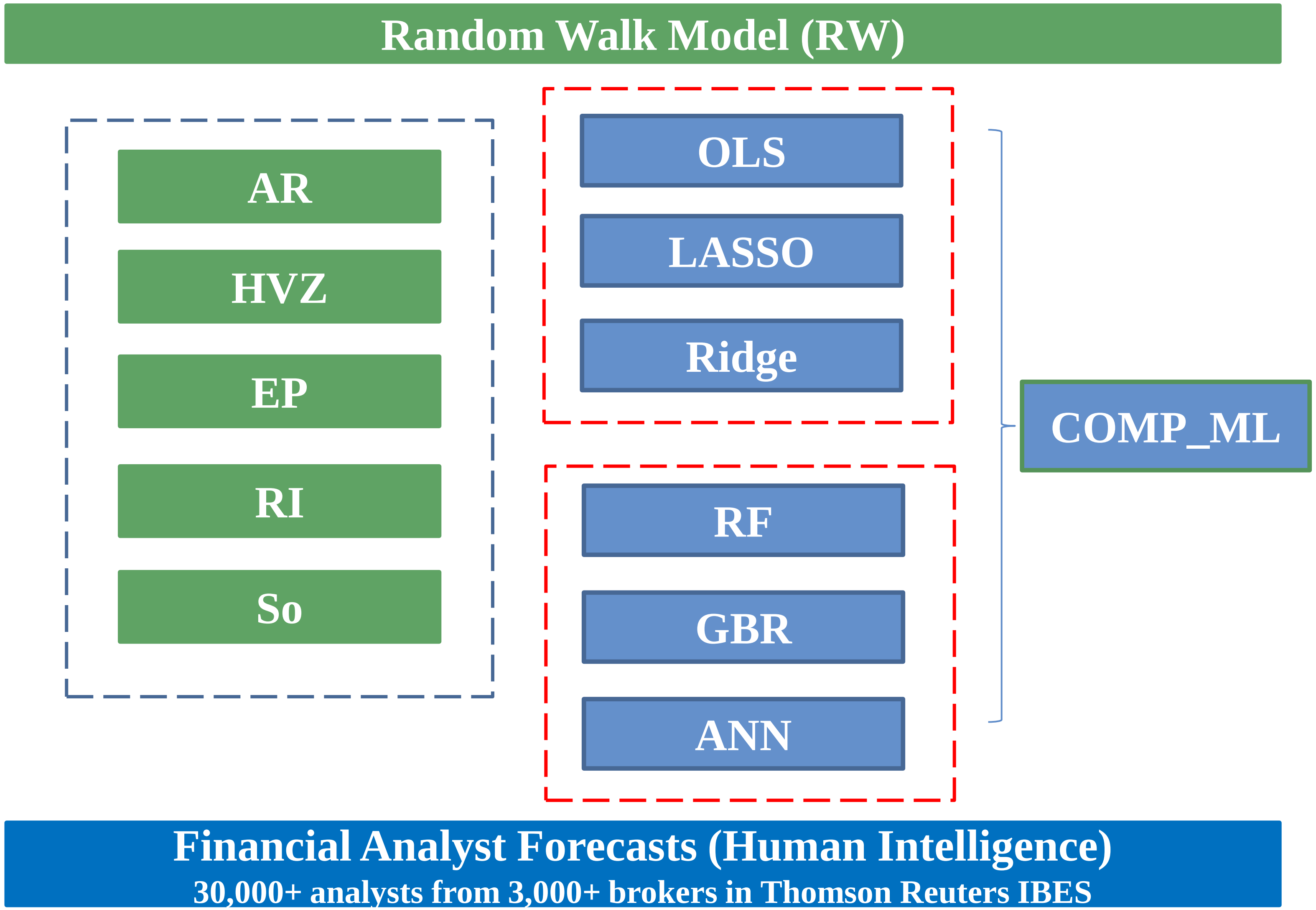
Income statement items (# = 12):	
$SALE_t$	Sales (sale)
$COGS_t$	Cost of goods sold (cogs)
$XSGA_t$	Selling, general, and administrative expenses (xsga)
XAD_t	Advertising expense (xad)
XRD_t	Research and development (R&D) expense (xrd)
DP_t	Depreciation and amortization (dp)
$XINT_t$	Interest and related expense (xint)
$NOPIO_t$	Non-operating income (expense) – other (nopio)
TXT_t	Income taxes (txt)
$XIDO_t$	Extraordinary items and discontinued operations (xido)
E_t	Earnings (ib - spi)
DVC_t	Common dividend (dvc)

Balance sheet items (# = 15):	
CHE_t	Cash and short-term investments (che)
$INVT_t$	Inventories (invt)
$RECT_t$	Receivables (rect)
ACT_t	Total current assets (act)
$PPENT_t$	Property, plant, and equipment – Net (ppent)
$IVAO_t$	Investments and advances – other (ivao)
$INTAN_t$	Intangible assets (intan)
AT_t	Total assets (at)
AP_t	Accounts payable (ap)
DLC_t	Debt in current liabilities (dlc)
TXP_t	Income taxes payable (txp)
LCT_t	Total current liabilities (lct)
$DLTT_t$	Long-term debt (dltt)
LT_t	Total liabilities (lt)
CEQ_t	Common/Ordinary equity (ceq)

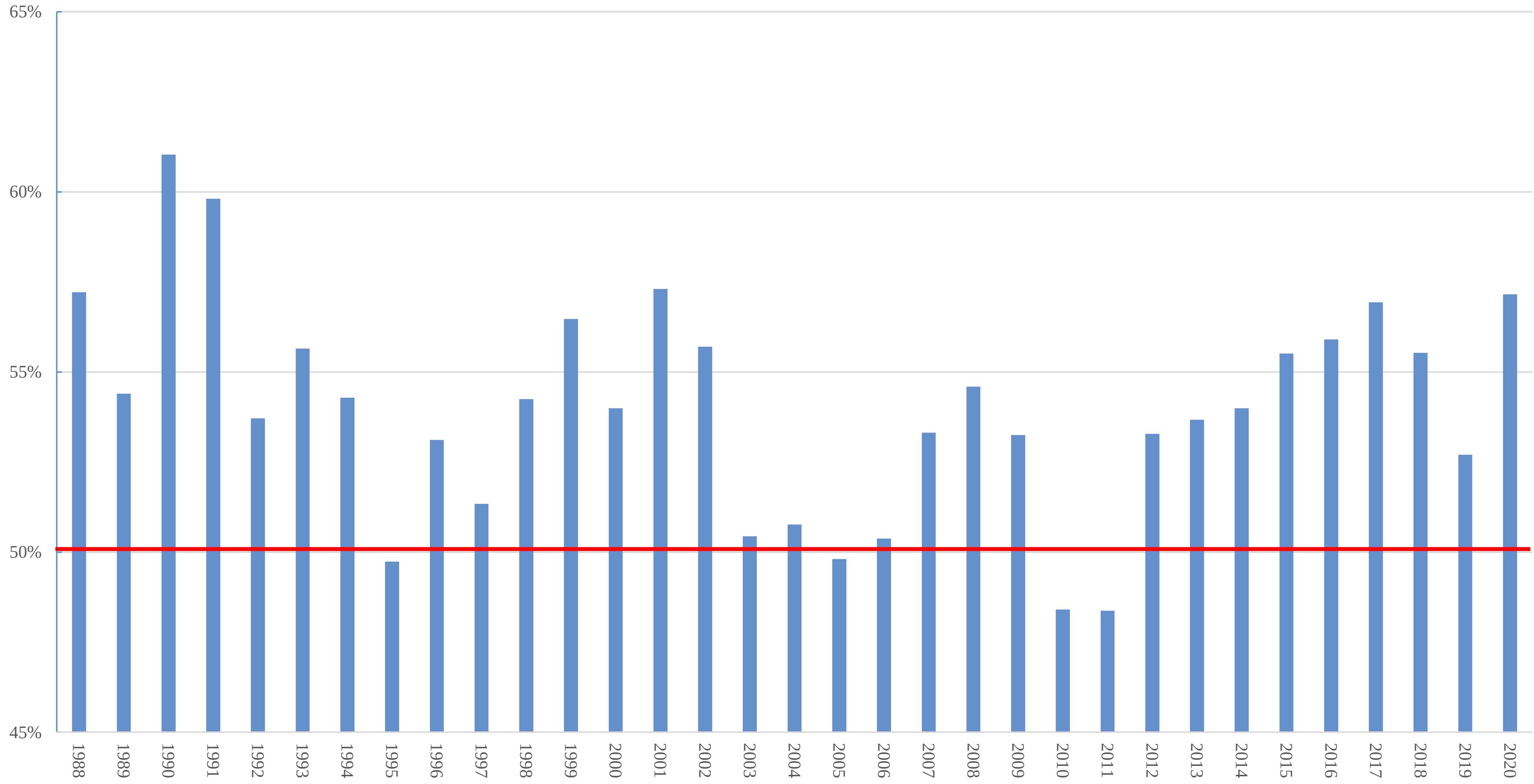
Cash flow statement items (# = 1):	
CFO_t	Cash flow from operating activities (oancf - xidoc); if missing, it is computed using the balance sheet approach (ib - accruals)

First-order differences of the above 28 items (# = 28):	
$\Delta CHE_t \sim \Delta CFO_t$	Computed as the corresponding item in year t less the same item in year t - 1

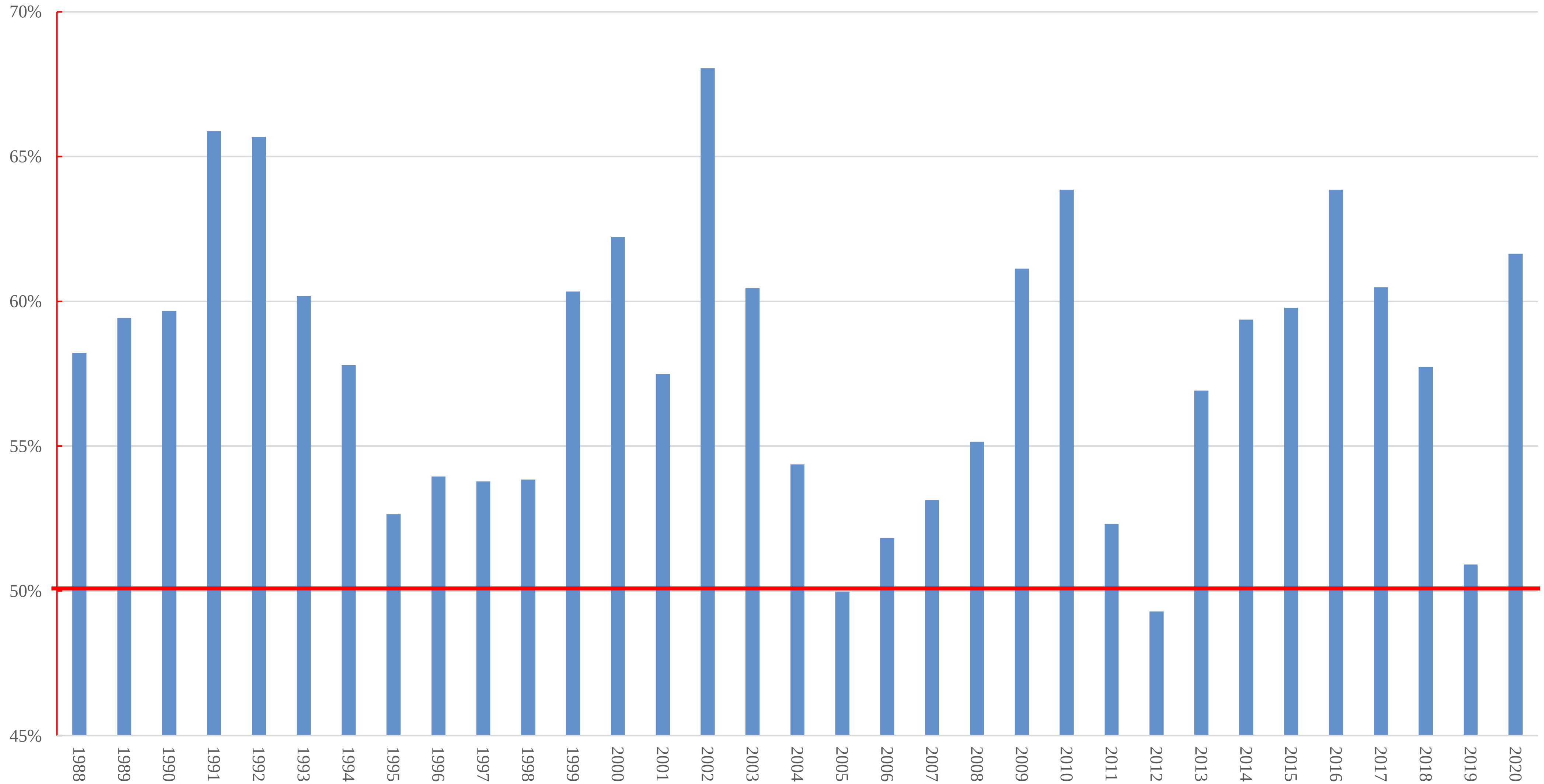
Model Selection



AI vs. Human (One Year ahead Forecasts): Percentage of firms where AI is more accurate



AI vs. Human (Two Years ahead Forecasts): Percentage of firms where AI is more accurate



AI vs. Human (Three Years ahead Forecasts): Percentage of firms where AI is more accurate

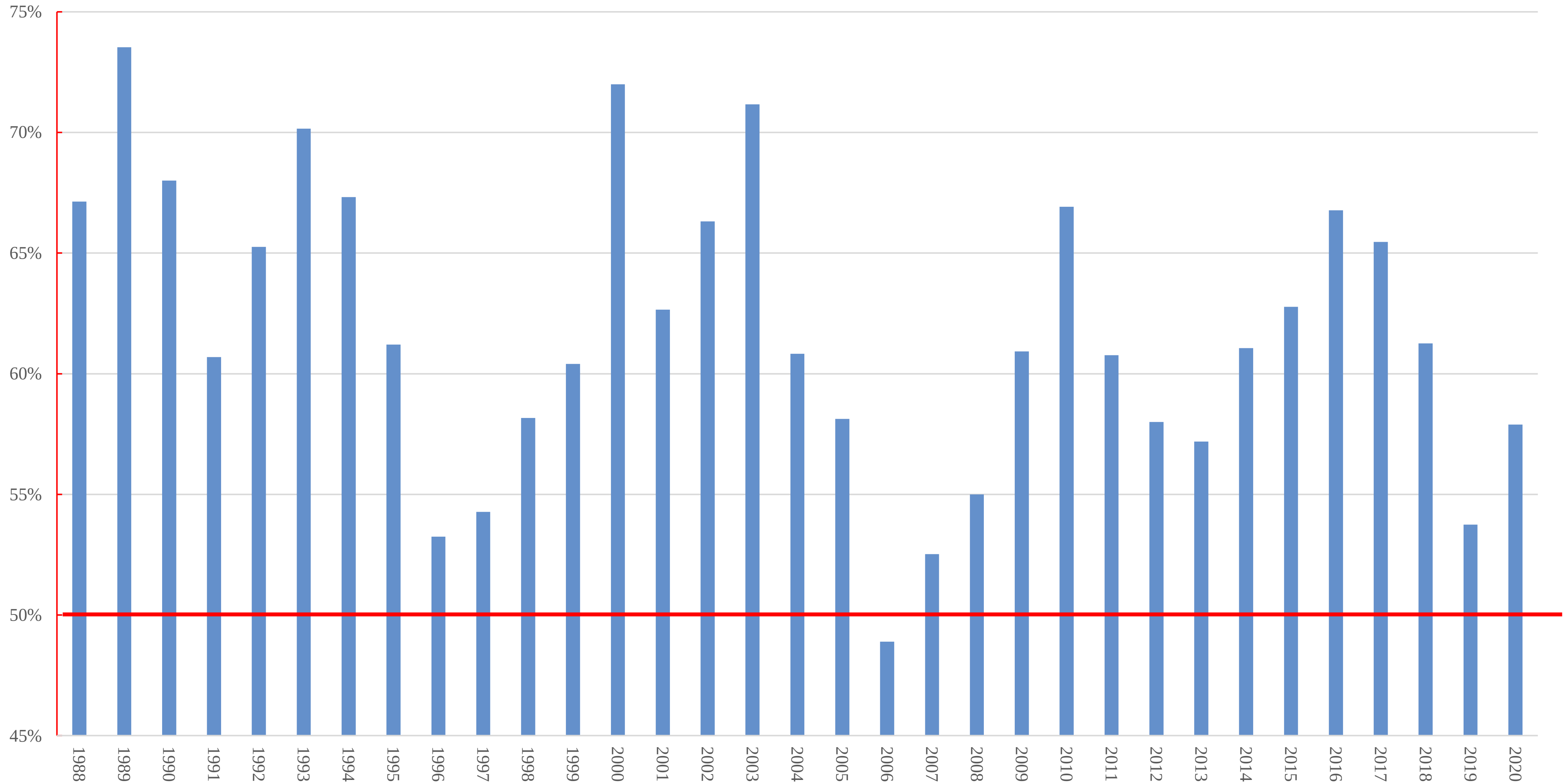


Table 7: Portfolio analysis of the new information uncovered using the machine learning models

Panel A: Equal-weighted portfolios

	OLS	LASSO	Ridge	RF	GBR	ANN	COMP_LR	COMP_NL	COMP_ML
Mean Return	0.6185	0.6262	0.6346	0.5962	0.6795	0.7185	0.6402	0.7203	0.7720
	(8.65)	(8.89)	(8.85)	(7.49)	(8.73)	(8.12)	(9.29)	(8.05)	(9.50)
CAPM Alpha	0.6817	0.6856	0.6989	0.6328	0.7110	0.7784	0.7022	0.7695	0.8372
	(9.96)	(10.46)	(10.48)	(7.82)	(9.07)	(8.89)	(10.87)	(8.78)	(10.73)
FF3 Alpha	0.6538	0.6597	0.6758	0.6062	0.6733	0.7247	0.6761	0.7279	0.8033
	(9.71)	(9.88)	(10.18)	(8.54)	(9.90)	(9.63)	(10.46)	(9.61)	(11.39)
Carhart4 Alpha	0.5938	0.5921	0.6178	0.5166	0.5934	0.6558	0.6137	0.6448	0.7134
	(9.08)	(9.03)	(9.49)	(7.29)	(8.57)	(8.50)	(9.66)	(8.35)	(10.23)
FF5 Alpha	0.5371	0.5488	0.5655	0.4312	0.4828	0.5286	0.5613	0.5143	0.6096
	(7.96)	(8.21)	(8.48)	(5.97)	(7.08)	(7.18)	(8.64)	(6.63)	(8.59)

Panel B: Value-weighted portfolios

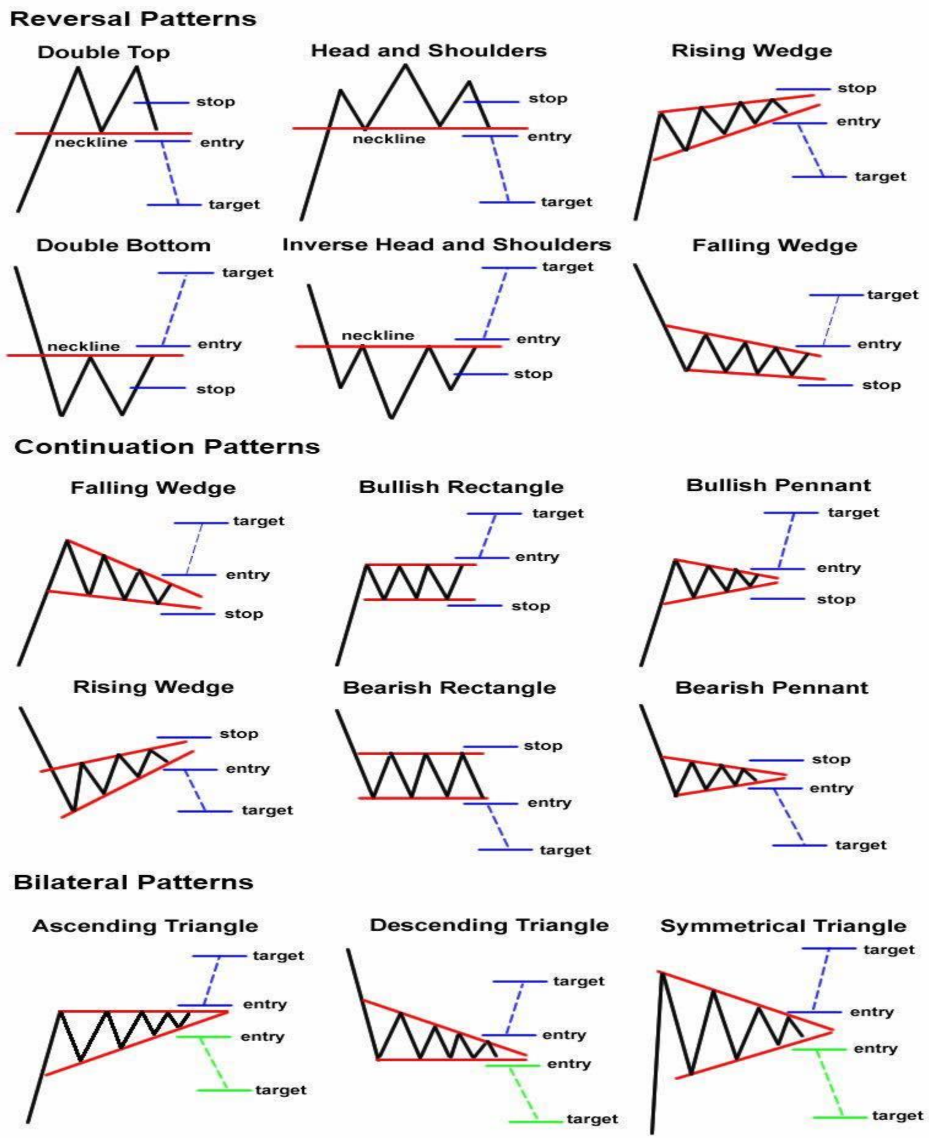
	OLS	LASSO	Ridge	RF	GBR	ANN	COMP_LR	COMP_NL	COMP_ML
Mean Return	0.2239	0.2484	0.2674	0.3177	0.4163	0.4747	0.2677	0.4568	0.3831
	(1.99)	(2.19)	(2.27)	(2.74)	(3.50)	(4.08)	(2.29)	(3.74)	(3.60)
CAPM Alpha	0.3571	0.3778	0.3969	0.3775	0.4797	0.5914	0.3954	0.5490	0.4884
	(3.30)	(3.57)	(3.53)	(3.05)	(4.01)	(5.07)	(3.58)	(4.34)	(4.66)
FF3 Alpha	0.3237	0.3552	0.3667	0.4478	0.5505	0.6325	0.3663	0.6217	0.5289
	(3.34)	(3.53)	(3.54)	(3.75)	(4.60)	(5.52)	(3.65)	(5.19)	(5.15)
Carhart4 Alpha	0.2829	0.2999	0.3320	0.3081	0.4316	0.5605	0.3247	0.4768	0.4558
	(3.08)	(3.06)	(3.41)	(3.07)	(3.70)	(4.70)	(3.37)	(4.49)	(4.23)
FF5 Alpha	0.1222	0.1205	0.1634	0.2810	0.4142	0.4358	0.1575	0.4119	0.3715
	(1.42)	(1.40)	(1.90)	(2.57)	(3.80)	(4.40)	(1.85)	(3.54)	(3.89)

AI-based Technical Analysis

- **Investors have used price charts and price patterns as tools for predicting future price movements for as long as there have been financial markets.**
  - **Prices reflect supply and demand forces**
  - **Price/volume patterns may shed light on whether prevailing trends will persist or reverse**
- **Price charts are often very subtle**
- **AI has outperformed humans in image recognition for several years**
  https://www.forbes.com/sites/michaelthomsen/2015/02/19/microsofts-deep-learning-project-outperforms-humans-in-image-recognition/?sh=5c12dae1740b
- **Can AI excel in technical analysis and help us predict stock returns?**

Price Charts

Technical Analysis & Price Patterns



Jiang, Kelly and Xiu (2023)
Research Design: Data and Feature Selection

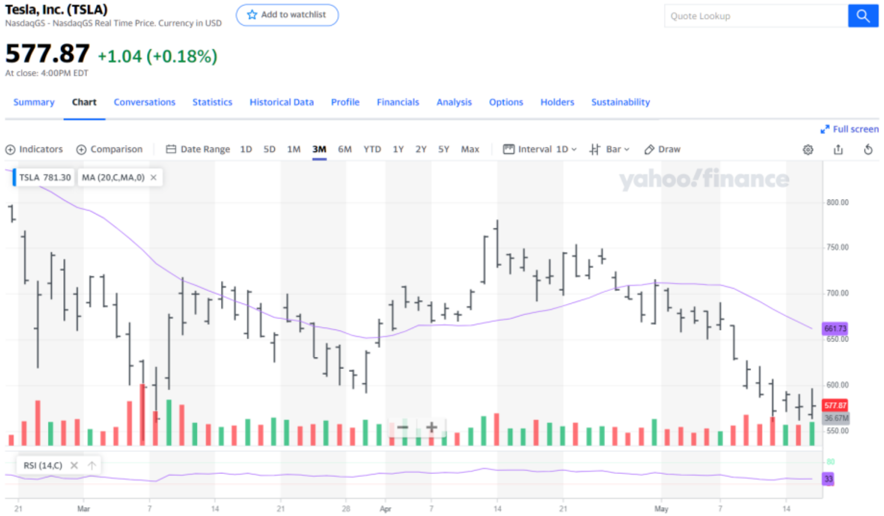
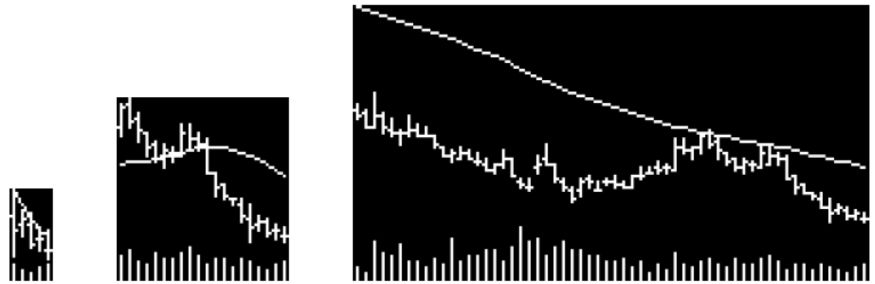


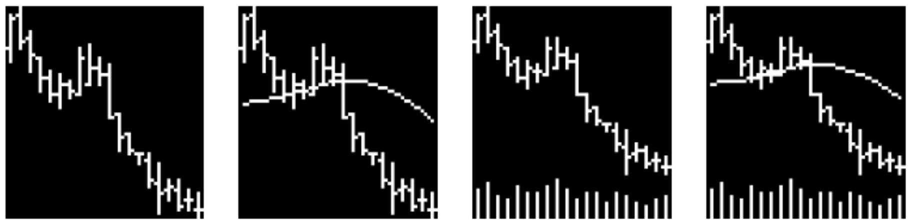
Figure 4: Generated OHLC Images with Volume Bar and Moving Average Line



(a) 5 (b) 20 (c) 60

Note: Market data images for 5, 20, and 60 days of data.

Figure 3: Examples of 20-day Image under Different Settings



(a) w/o VB, w/o MA (b) w/o VB, w/ MA (c) w/ VB, w/o MA (d) w/ VB, w/ MA

Note: From left to right are 20-day images (a) without volume bar and moving average line, (b) without volume bar but with moving average line, (c) with volume bar but without moving average line, and (d) with volume bar and moving average line.

Jiang et al. (2023)
Research Design

- **Sample: NYSE, AMEX, and NASDAQ**
- **Sample period: 1993-2019**
- **Training & Validation:**
  - **1993 to 1999**
  - **70% training & 30% for validation (randomly)**
- **Test sample: 2000-2019**
- **Target variable: y=1 if subsequent return is positive and y=0 otherwise**

Out-of-Sample Classification Accuracy

Table 2: Out-of-Sample Classification Accuracy

	Return horizon			
	20-day		60-day	
Image size	Acc.	Corr.	Acc.	Corr.
5-day	52.1%	3.2%	52.5%	2.0%
20-day	52.5%	3.2%	52.9%	2.6%
60-day	52.5%	3.1%	53.5%	3.1%
MOM	52.2%	1.9%	52.2%	1.7%
STR	50.4%	1.4%	49.7%	1.2%
WSTR	51.1%	2.8%	50.6%	2.6%

Note: The table reports out-of-sample forecast performance for image-based CNN models and benchmark signals. We calculate classification accuracy and correlation cross-sectionally each period then report time series averages over each period in the test sample.

Short-horizon Portfolio Analysis

Table 6: Short-horizon (One Week) Portfolio Performance

	Equal Weight											
	I5/R5		I20/R5		I60/R5		MOM/R5		STR/R5		WSTR/R5	
	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Low	-0.29	-2.04	-0.34	-2.14	-0.22	-1.21	0.14	0.41	-0.02	-0.10	-0.09	-0.41
2	-0.07	-0.44	-0.06	-0.36	-0.01	-0.07	0.08	0.36	0.04	0.22	0.02	0.12
3	0.00	0.00	0.01	0.06	0.06	0.30	0.08	0.37	0.06	0.41	0.05	0.32
4	0.04	0.22	0.07	0.37	0.08	0.40	0.07	0.41	0.08	0.49	0.06	0.41
5	0.08	0.43	0.10	0.51	0.11	0.60	0.07	0.44	0.08	0.50	0.07	0.42
6	0.10	0.51	0.13	0.66	0.13	0.72	0.09	0.57	0.09	0.53	0.08	0.49
7	0.15	0.73	0.16	0.84	0.15	0.81	0.10	0.66	0.09	0.50	0.11	0.62
8	0.20	0.96	0.20	1.01	0.18	0.97	0.12	0.77	0.10	0.51	0.12	0.62
9	0.28	1.38	0.26	1.31	0.21	1.15	0.14	0.82	0.14	0.62	0.17	0.74
High	0.53	2.79	0.50	2.67	0.32	1.78	0.16	0.74	0.38	1.16	0.45	1.53
H-L	0.82***	6.99	0.85***	6.89	0.55***	5.17	0.02	0.07	0.40***	1.78	0.54***	2.88
Turnover	847%		820%		764%		130%		358%		725%	

	Value Weight											
	I5/R5		I20/R5		I60/R5		MOM/R5		STR/R5		WSTR/R5	
	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR	Ret	SR
Low	-0.06	-0.37	-0.06	-0.37	-0.05	-0.28	0.01	0.02	0.02	0.09	-0.04	-0.16
2	-0.01	-0.09	0.01	0.06	0.00	0.02	-0.01	-0.02	0.02	0.10	0.00	0.03
3	0.03	0.19	0.02	0.11	0.02	0.13	0.04	0.17	0.03	0.20	0.04	0.21
4	0.02	0.11	0.02	0.14	0.01	0.06	0.05	0.23	0.07	0.41	0.04	0.22
5	0.06	0.33	0.04	0.22	0.03	0.15	0.05	0.28	0.07	0.43	0.05	0.33
6	0.05	0.28	0.06	0.32	0.05	0.29	0.05	0.31	0.08	0.44	0.08	0.47
7	0.09	0.51	0.08	0.47	0.06	0.34	0.06	0.38	0.07	0.39	0.09	0.52
8	0.11	0.57	0.09	0.51	0.08	0.45	0.08	0.50	0.11	0.53	0.13	0.64
9	0.13	0.67	0.11	0.57	0.10	0.53	0.09	0.53	0.11	0.43	0.16	0.65
High	0.19	0.86	0.17	0.86	0.13	0.73	0.13	0.59	0.15	0.42	0.17	0.54
H-L	0.25***	1.63	0.24***	1.69	0.19***	1.57	0.13	0.36	0.13**	0.45	0.21***	0.78
Turnover	979%		869%		895%		121%		430%		840%	

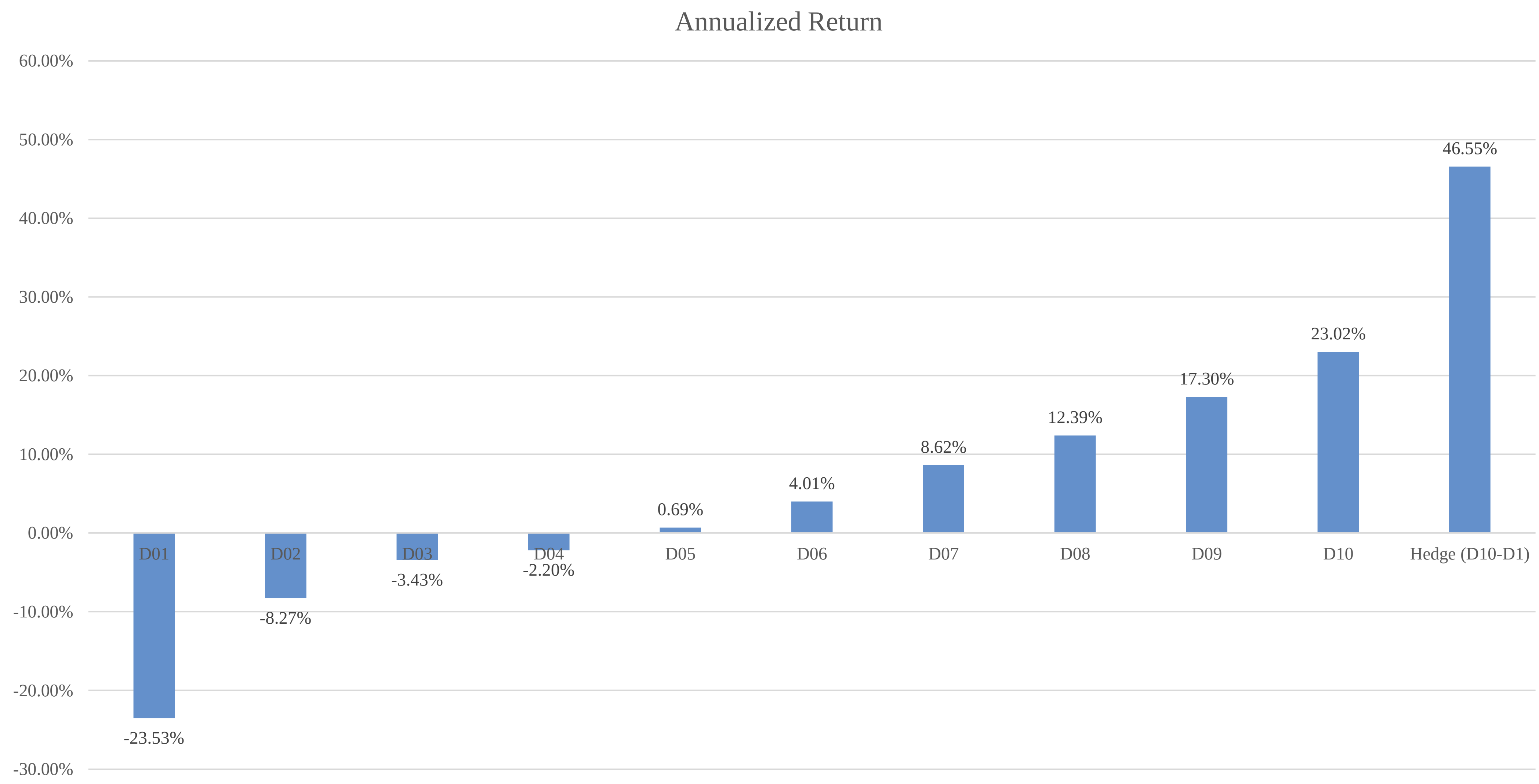
Note: Performance of equal-weighted (top panel) and value-weighted (bottom panel) decile portfolios sorted on out-of-sample predicted up probability. Each panel reports the average holding period return and annualized Sharpe ratios. Average returns accompanied by ***,**,* are significant at the 1%, 5% and 10% significance level, respectively. We also report monthly turnover of each strategy.

AI based Technical Analysis on China Markets: Similar K-Lines

AI based Technical Analysis on China Markets

- **For at the end of week *t*, for stock *i*, identify 5,000 cases in the training dataset with the most similar X-day price patterns (i.e. KNN)**
- **Obtain return prediction for $S_{i,t}$ based on the distribution of the subsequent Y-day stock returns of the 5,000 cases**
- **Repeat the above process for all stock-week pairs**
- **Sort all stocks in the CSI800 universe into 10 deciles based on the above return prediction and hold the portfolio for one week.**
- **Repeat the above analysis at the end of next week.**

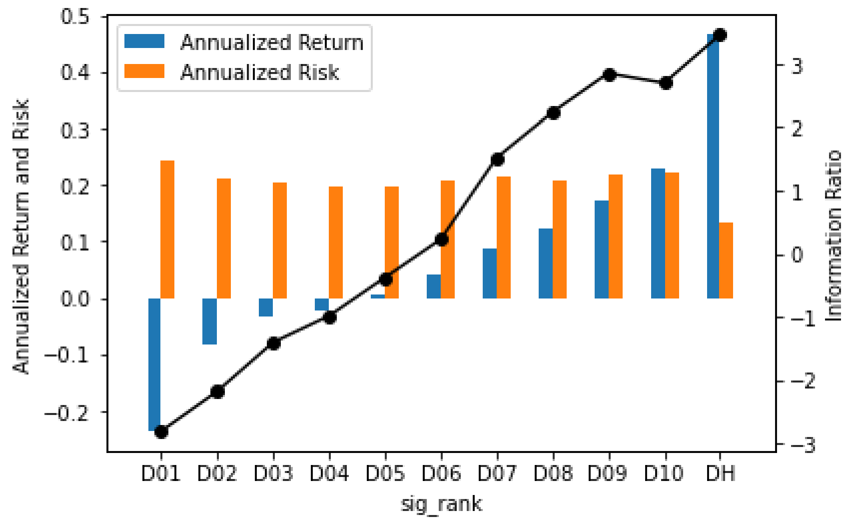
Annualized Returns to Portfolios sorted on Predicted Return based on Similar Price Charts

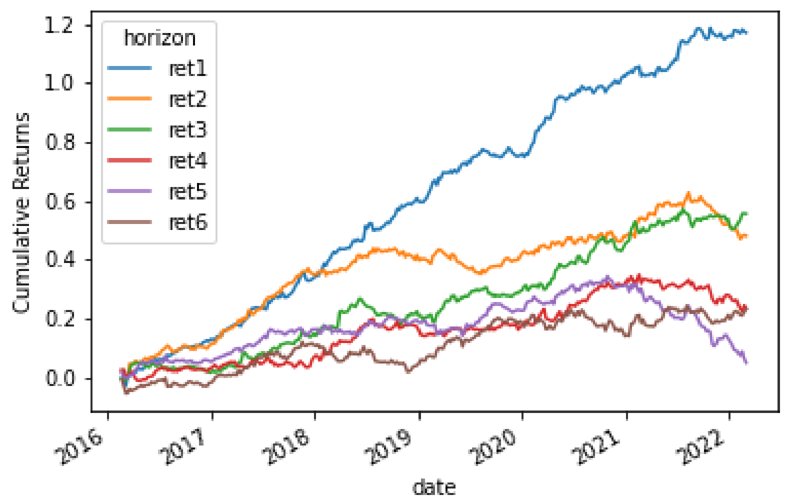


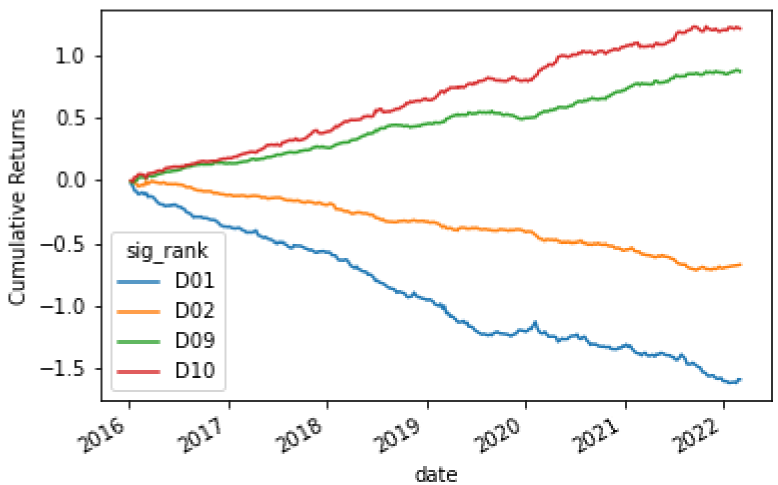
Sample Signal:
Pattern Recognition with Machine Learning

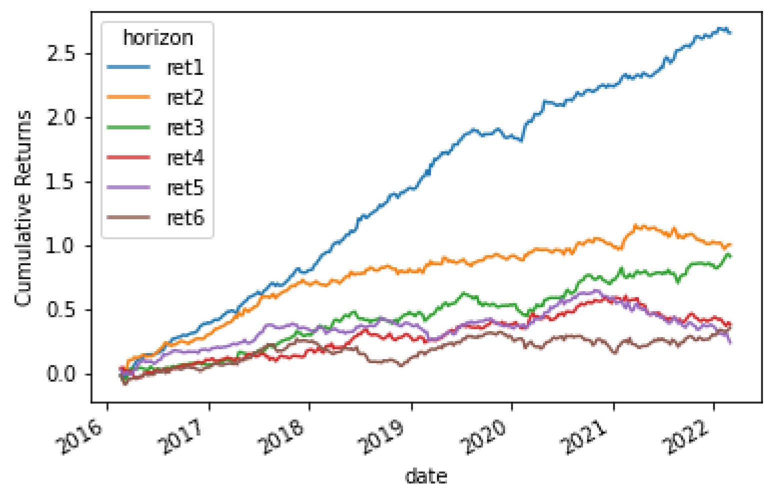
sig_rank	Annualized Return	Annualized Risk	Sharpe Ratio	Annualized Active Return	Annualized Active Risk	Information Ratio	Max Drawdown (Raw)	Max Drawdown (Active)	Turnover (annualized)
D01	-23.53%	24.18%	-0.973	-26.39%	9.42%	-2.801	77.58%	80.43%	77.14
D02	-8.27%	21.15%	-0.391	-11.13%	5.12%	-2.173	50.49%	51.48%	89.21
D03	-3.43%	20.52%	-0.167	-6.28%	4.48%	-1.401	39.80%	34.84%	90.38
D04	-2.20%	19.67%	-0.112	-5.05%	5.14%	-0.983	38.83%	32.93%	90.14
D05	0.69%	19.69%	0.035	-2.17%	5.75%	-0.377	44.74%	26.36%	85.62
D06	4.01%	20.69%	0.194	1.15%	5.06%	0.228	41.44%	14.27%	89.46
D07	8.62%	21.47%	0.401	5.76%	3.80%	1.517	30.10%	3.20%	91.72
D08	12.39%	20.92%	0.593	9.54%	4.27%	2.234	26.03%	2.80%	91.03
D09	17.30%	21.69%	0.798	14.45%	5.07%	2.852	22.84%	6.29%	89.83
D10	23.02%	22.34%	1.030	20.16%	7.47%	2.698	20.50%	4.06%	84.69
DH	46.55%	13.45%	3.462	46.55%	13.45%	3.462	9.43%	9.43%	80.92

Sample Signal:
Pattern Recognition with Machine Learning









Textual Data and NLP

- **Much of the data produced today is text from various sources such as web, social media, newswire, emails, regulatory documents…**

- **How do investors make sense of text data?**

- **Natural Language Processing (NLP) helps to convert texts (unstructured) into an easier to use format (structured).**

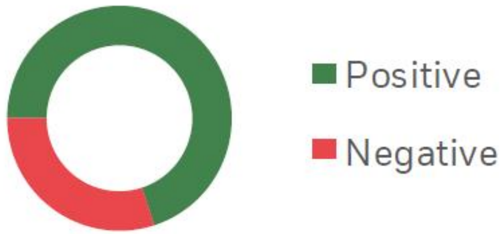
Use-case: NLP Analysis on Earnings Guidance by Blackrock

Example of how we analyze large data sets to identify signals

Using text analysis techniques to anticipate future changes to company earnings guidance

Analyze

Use technology to analyze over 5,000 earnings call transcripts every quarter and more than 6,000 broker reports every day



Measure

Transform unstructured text into proprietary measures of trending analyst sentiment

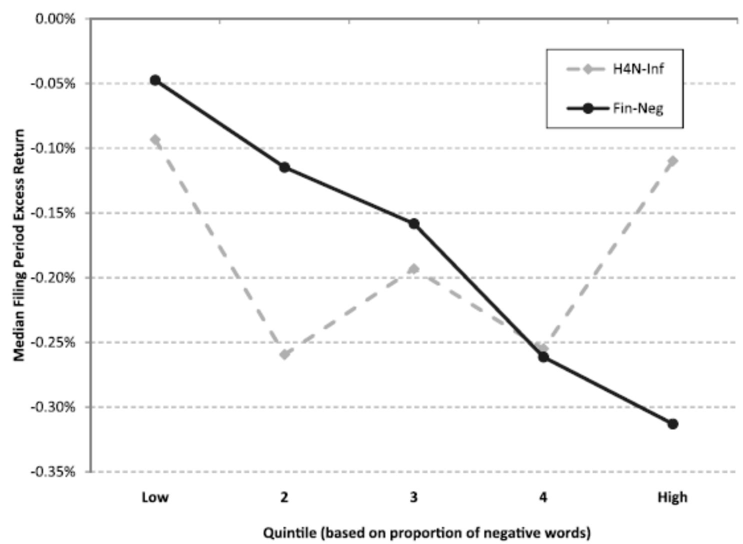
Traditional approach:
Individual reports read by hand – or await analyst revision to occur

NLP and Sentiment Analysis

- **Data Preprocessing**
  - **Tokenization: covert sentences to words**
  - **Remove stop words-frequent words such as “the”, “is”, etc.**
  - **Stemming and lemmatization: reduce words to its root (playing, plays, played=> play)**
- **Sentiment Analysis**
  - **Dictionary based approach: positive/negative words: https://sraf.nd.edu/textual-analysis/resources/**
  - **Machine learning approach:**
    - Feature extraction: mapping text to real value vector (Bag of Words and Word2vec etc.)
    - Train a machine learning algorithm

Dictionary based measure of sentiment

- **Harvard General Inquirer list: http://www.wjh.harvard.edu/~inquirer**
- **Loughran and McDonald (2011)**
  - **A word list developed for psychology and sociology may not translates well into business, for example, t*ax*, *cost*, *capital*, *board*, *liability*, *foreign*, and *vice* are negative on the Harvard list**
  - **Create a list of 2,354 words that typically have negative implications in a financial sense, and a list of 354 positive words (https://sraf.nd.edu/textual-analysis/resources/)**



FinBert by Huang, Wang and Yang (2020)

- **BERT (Bidirectional Encoder Representations from Transformers), Google's state-of-the-art language model for NLP, which learn the language model by:**
  - **Masked Language Modeling (LM): randomly mask 15% of the words with a [MASK] token, and then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked words in the sequence**
  - **Next Sentence Prediction (NSP): the model receives pairs of sentences as input and learns to predict whether the 2nd sentence in the pair is the subsequent sentence in the original document.**

BERT Fine-Tuning for Specific Tasks

- **Google pre-trained two BERT models using general text copus from Wikipedia and BooksCorpus with a total of 3.3 billion word tokens:**

$BERT_{BASE}$	$BERT_{LARGE}$
Layers = 12	Layers = 24
Hidden size = 768	Hidden size = 1024
self-Attention heads = 12	self-Attention heads = 16
Total parameters = 110M	Total parameters = 340M

- **Using transfer learning, users can fine-tune the pre-trained model for specific tasks such as sentiment analysis, question-answering tasks, and named entity recognition etc.**
- **<u>Sentiment analysis:</u> adding a classification layer on top of the transformer output to predict sentiment labels (by human), just like the Next Sentence classification**
- **Huang et al. (2020)**
    - **Pre-train the FinBERT based on the pretrained BERT by google using financial text in 10-K, 10-Q, Earnings conference call and Analyst Report**
    - **Fine-tune the FinBERT model for sentiment classification using a sample of 10,000 pre-labeled sentences from financial text**

Performance of Sentiment Score of FinBERT

- Sentiment classification accuracy
  - FinBERT: 88.4%, Loughran and McDonald: 61.7%, BERT: 85.5
- FinBERT based sentiment score has higher association with market reaction to conference calls and abnormal trading volume
- FinBERT based sentiment score also predict future earnings better than the sentiment score based on the LM dictionary

Panel A: Regression of cumulative abnormal return on textual sentiments

	(1)	(2)	(3)	(4)	(5)
Dependent Variable			*CAR*		
$Tone_{FinBERT}$	0.734***				
	(15.01)				
$Tone_{BERT}$		0.709***			
		(15.08)			
$Tone_{LM}$			0.464***		
			(9.77)		
$Tone_{NB}$				0.369***	
				(8.10)	
$Tone_{W2V}$					0.175***
					(3.86)

AI Reads Chinese Analyst Reports

- Use a dictionary of positive and negative words in Chinese
- Computer program reads the abstract of analyst reports and assign a sentiment score based on the ratio of positive vs. negative words

One of the Top 5 most positive report in 2022

中信证券2021年业绩快报点评：21年业绩大超预期，维持行业首推

本报告导读：得益于资本市场的蓬勃发展，2021年公司各项业务均衡发展，稳步增长。全年业绩增速超预期。我们维持目标价36.77元，维持“增持”评级。

投资要点：维持“增持”评级，维持目标价36.77元，对应22年17.1xP/E：受益于财富管理需求爆发及注册制改革，中信证券投行及资产管理超预期带来整体盈利增速超预期。故我们上调21/22/23年利润为231.52/277.65/325.08亿元（调整前为206.66/228.44/272.63亿元），对应EPS1.79/2.15/2.51元，我们维持增持评级，维持目标价36.77元，对应22年17.1倍P/E。

受益于财富管理与注册制改革，业绩增长超预期：根据公司披露的业绩快报，中信证券2021年营业收入同比+40.80%；归母净利润同比+54.20%，盈利增速大超预期。我们认为居民财富管理需求爆发带来的财富管理产业链的收入及注册制改革带来的投行业务的高增长是盈利超预期的主要原因。在2021年公司机构业务受制于资本金约束，增长收到影响的情况下，财富管理和投行业务成为盈利重要增长点。

公司配股将近，资本金得到补充后机构业务有望发力。未来投资者机构化趋势将带来机构业务需求的快速提升，我们认为2022年机构业务将迎来高速增长。随着配股的落地，公司资本金将得到大幅补充，NSFR/LCR流动性约束得到满足，为后续机构业务的全面发力打下了坚实基础，在同业受制于资本约束的背景下，公司在机构业务上的市场份额有望进一步提升。

催化剂：市场交投活跃度提升，机构客户需求旺盛

风险提示：权益市场大幅波动；机构业务开展低于预期

One of the Top 5 most negative report in 2022

万科A2021年经营点评：严推盘，扩权益，底线思维提升经营质量

核心观点：

低景气度维持谨慎的推盘策略。根据2021年12月经营公告，万科12月实现合同销售金额635.6亿元，同比下降37.4%。累计来看，21年全年销售金额6277.8亿元，同比下降10.8%，累计销售面积3807.8万方，同比下降18.4%，销售均价16487元/平，同比增长9.3%。下半年基本面下行压力加大，公司采取谨慎的推盘策略。

21年拿地金额行业前二，权益投资增长10%。2021年全年，万科共计获取152个项目，拿地金额1909亿元，拿地面积2823万方，金额及面积口径的拿地力度分别为30%、74%，与20年略有提升。2021万科全口径拿地金额行业前二，权益投资1569亿元，同比增长10%，权益比例从20年的69%提升至82%。

投资聚焦核心城市，上海区域占比提升。万科21年无新进入城市，布局10年以上城市投资金额占比为72%，与过去5年平均水平（71%）基本一致。上海区域投资比例40%，较20年上升4pct，南方维持25%，中西部维持16%，西北上升至8%，北京区域降至13%。21年共获取36宗两集中地块，总地价457亿元，占全年拿地金额的24%。

盈利预测与投资建议。万科整体经营稳健，低景气度谨慎推盘，保持良好的投资纪律。行业资源向优质“经营绿档”公司倾斜，将促使公司主业增速提升。预计21-22年业绩361亿元、372亿元，同比分别-13%、+3.0%，对应6.8xPE、6.6xPE，维持合理价值26.42元/股，考虑到过去5年A股相对H股溢价15%，按照汇率及溢价计算，对应H股28.14港元/股，维持A\H股“买入”评级。

风险提示。行业景气度下行影响公司销售，结算规模不及预期。

AI Reads Chinese Analyst Reports (Dictionary approach)

- **Form ten decile portfolios based on the average analyst sentiment over the past 3 months.**

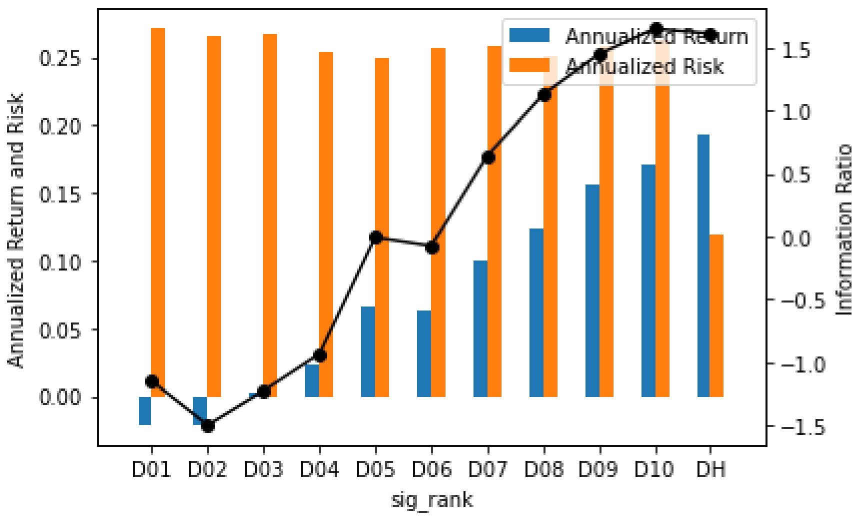
AI Reads Chinese Analyst Reports (Deep learning approach)

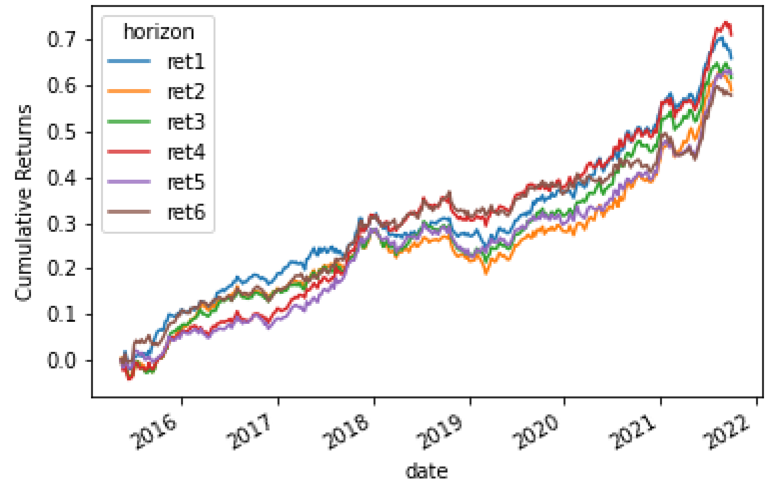
- **Form ten decile portfolios based on the average analyst sentiment over the past 3 months.**

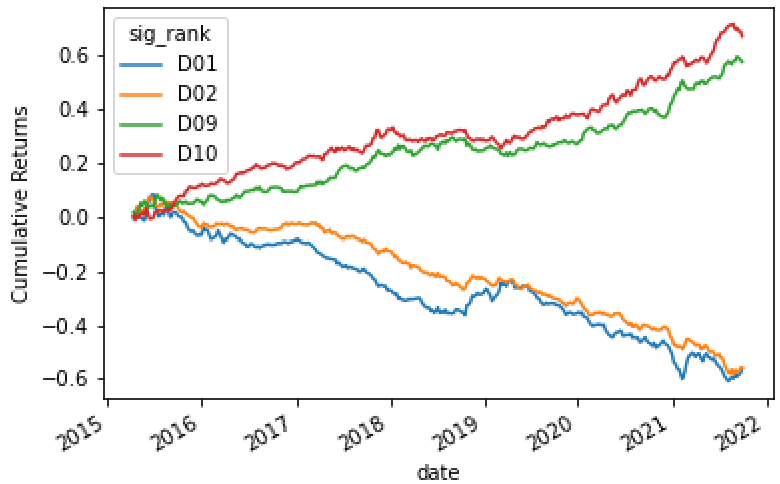
Sample Signal:
Sentiment of Analyst Reports – Deep Learning Approach

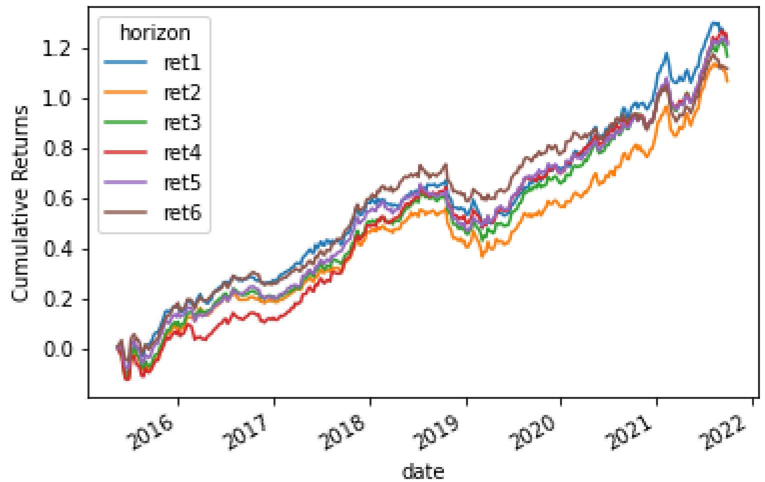
sig_rank	Annualized Return	Annualized Risk	Sharpe Ratio	Annualized Active Return	Annualized Active Risk	Information Ratio	Max Drawdown (Raw)	Max Drawdown (Active)	Turnover (annualized)
D01	-2.11%	27.10%	-0.078	-8.77%	7.63%	-1.150	70.05%	50.70%	15.12
D02	-2.06%	26.51%	-0.078	-8.73%	5.83%	-1.498	66.90%	48.79%	23.94
D03	0.31%	26.67%	0.012	-6.35%	5.18%	-1.226	66.31%	37.77%	28.50
D04	2.39%	25.39%	0.094	-4.28%	4.58%	-0.934	56.71%	25.18%	31.55
D05	6.64%	24.93%	0.267	-0.02%	4.69%	-0.005	46.15%	11.79%	33.70
D06	6.32%	25.62%	0.247	-0.35%	4.78%	-0.073	49.90%	9.33%	34.88
D07	10.00%	25.89%	0.386	3.34%	5.17%	0.645	49.73%	6.58%	34.69
D08	12.35%	25.10%	0.492	5.68%	5.03%	1.130	43.76%	4.96%	33.42
D09	15.65%	25.76%	0.608	8.99%	6.19%	1.453	45.77%	6.77%	29.90
D10	17.14%	26.25%	0.653	10.48%	6.33%	1.655	41.07%	7.67%	18.54
DH	19.25%	11.92%	1.616	19.25%	11.92%	1.616	17.75%	17.75%	16.83

Sample Signal:
Sentiment of Analyst Report – Deep Learning Approach





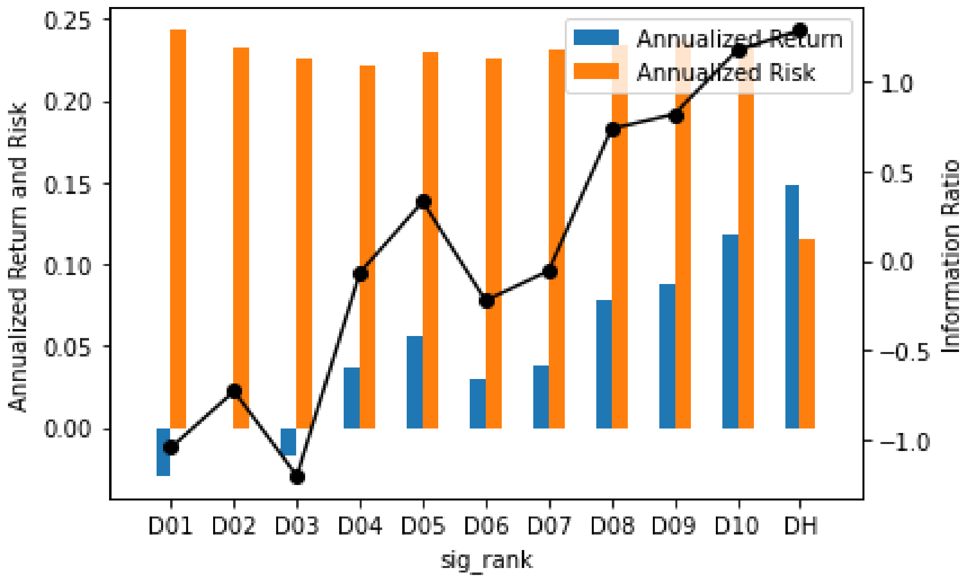


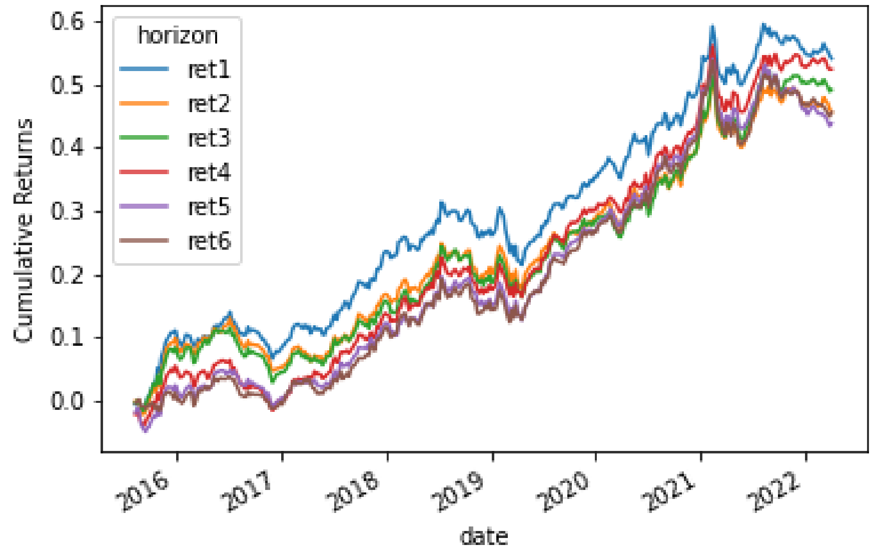


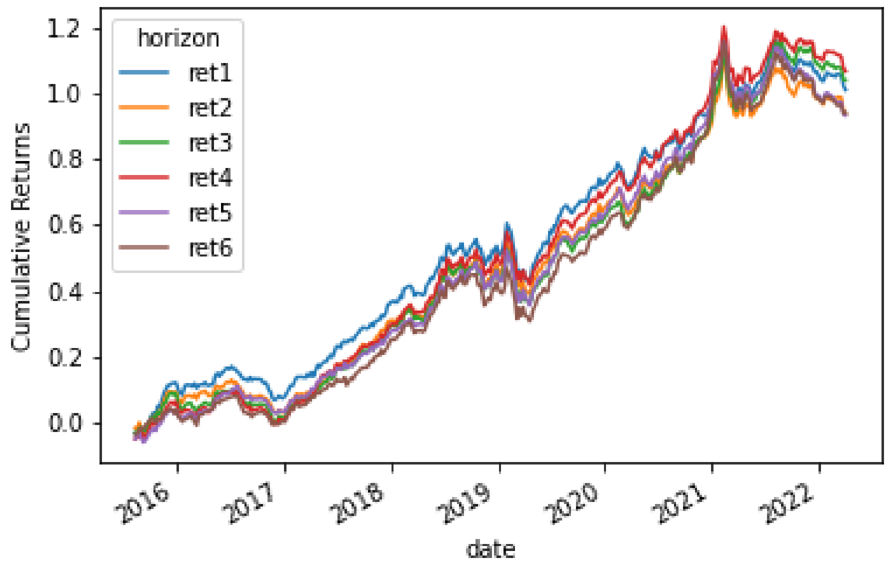
Sample Signal:
Sentiment of Financial News

	Annualized Return	Annualized Risk	Sharpe Ratio	Annualized Active Return	Annualized Active Risk	Information Ratio	Max Drawdown (Raw)	Max Drawdown (Active)	Turnover (annualized)
D01	-2.95%	24.31%	-0.121	-6.93%	6.64%	-1.043	57.36%	45.62%	14.82
D02	-0.03%	23.24%	-0.001	-4.01%	5.50%	-0.729	52.22%	33.82%	29.19
D03	-1.67%	22.56%	-0.074	-5.65%	4.70%	-1.202	49.53%	32.85%	36.01
D04	3.67%	22.21%	0.165	-0.31%	4.71%	-0.066	39.74%	16.73%	40.11
D05	5.55%	22.98%	0.241	1.57%	4.75%	0.330	42.20%	8.49%	41.82
D06	3.02%	22.57%	0.134	-0.96%	4.33%	-0.222	42.05%	12.73%	42.28
D07	3.74%	23.14%	0.162	-0.24%	4.28%	-0.056	44.28%	8.22%	41.50
D08	7.75%	23.46%	0.330	3.77%	5.11%	0.738	39.27%	4.83%	37.99
D09	8.83%	23.64%	0.374	4.85%	5.91%	0.820	37.46%	8.82%	31.43
D10	11.87%	23.46%	0.506	7.89%	6.70%	1.177	29.19%	9.67%	15.67
DH	14.82%	11.49%	1.289	14.82%	11.49%	1.289	16.42%	16.42%	15.25

Sample Signal:
Sentiment of Financial News

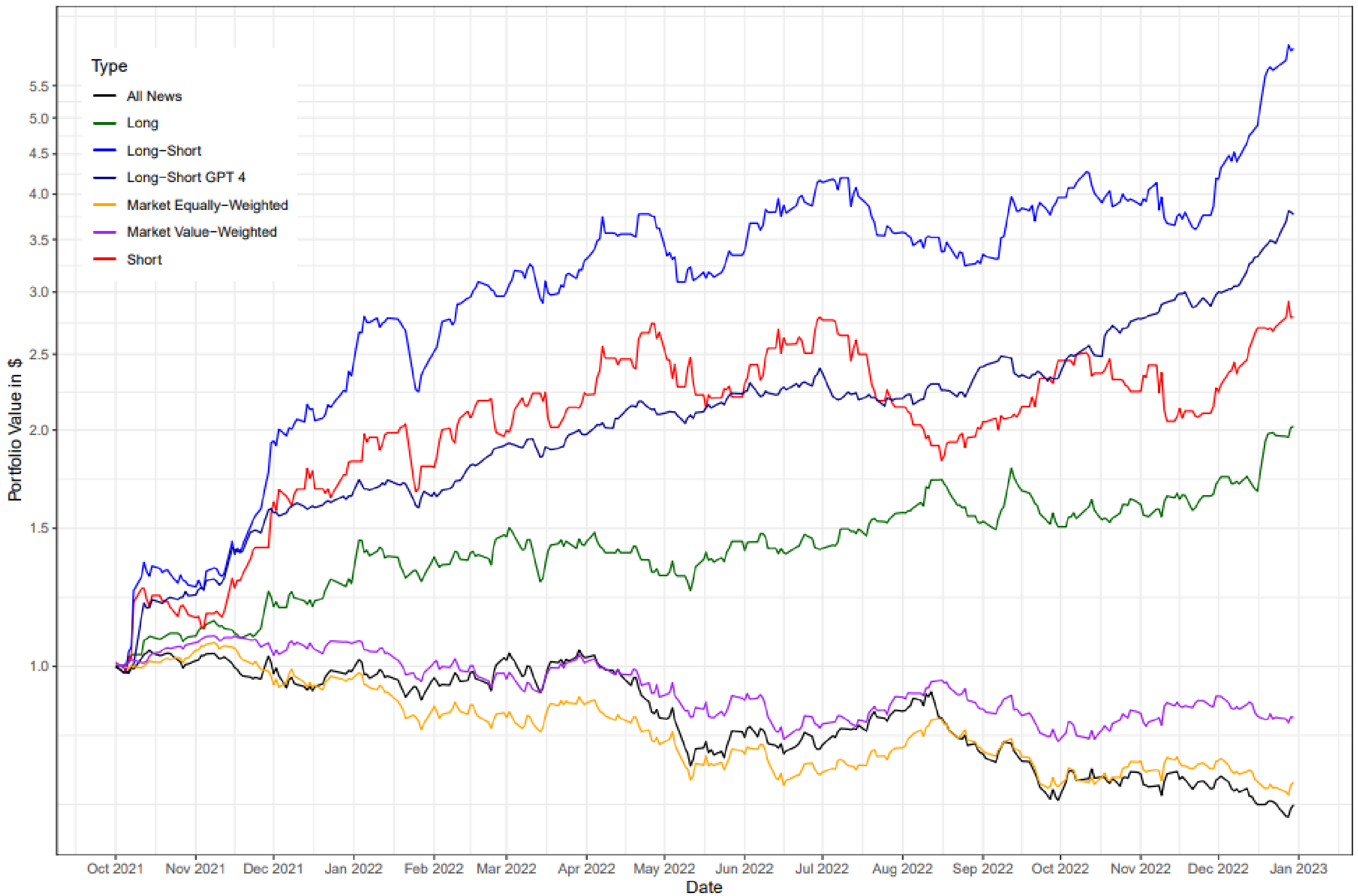






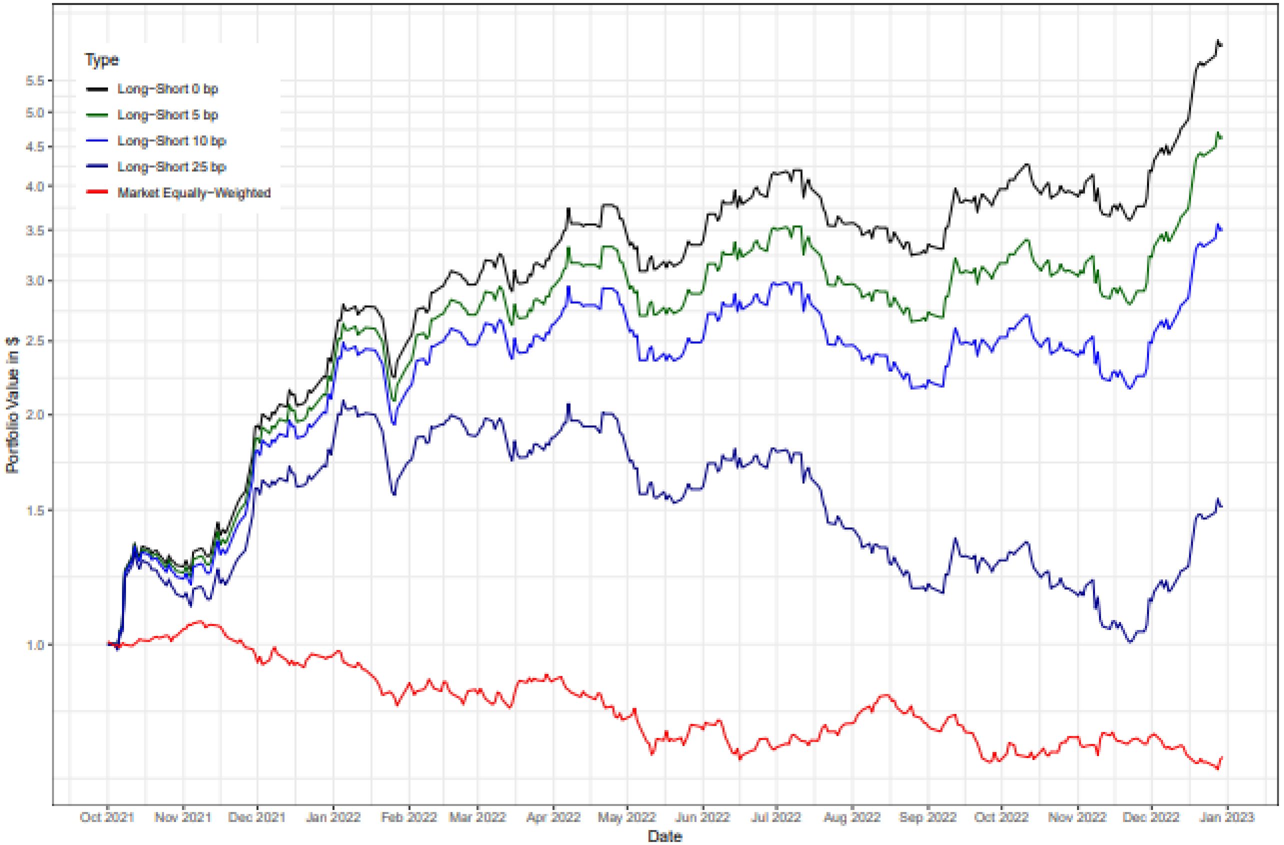
What about LLM or ChatGPT? (Lopez-Lira and Tang 2023)

Figure 1: Cumulative Returns of Investing $1 (Without Transaction Costs)



ChatGPT Strategy with Transaction Costs

Figure 2: Cumulative Returns of Investing $1 in the Long-Short Strategy for Different Transaction Costs



Four Strategies with Different Return-Risk Profiles

Summary of Backtest Performance (2017.01.04—2023.09.11）							
Strategies	CSI300	CSI500	CSI300 Enhanced Index	CSI500 Enhanced Index	Long Only Total Return	Market Neutral	Multi-Strategy
Annualized return	4.11%	0.91%	18.72%	26.34%	33.24%	15.07%	9.55%
Annualized risk	19.60%	21.44%	19.85%	22.29%	20.76%	6.10%	2.83%
Max drawdown	39.60%	41.01%	25.99%	28.41%	18.16%	4.70%	2.07%
Sharpe ratio	0.21	0.04	0.94	1.18	1.60	2.47	3.38
Annualized active return			14.61%	25.43%			
Annualized active risk			5.99%	7.17%			
Max active drawdown			6.08%	4.96%			
Information ratio			2.44	3.55			

***Daily Rebalance at next day open;**
***Transaction Cost: One-side 0.15%；**
***Index Future Hedging Cost：8%**

THANKS!