

資産運用におけるオルタナティブデータと 機械学習の応用事例とその課題



関西学院大学



Magne-Max Capital Management

岡田克彦

COVID-19 pandemic deals body blow to quant models, study shows

LONDON (Reuters) Oct 26. 2020 - The coronavirus pandemic has dealt a body blow to the quantitative model-based style of investing, with a majority of the firms using such strategies negatively impacted, a study by Refinitiv has found.

In a report, financial data provider **Refinitiv said 72% of such investors were hurt** by the pandemic. Some 12% declared their models obsolete and 15% were building new ones.

Machine-learning refers to the use of complicated mathematical models and algorithms based on historical data in order to make predictions without being explicitly programmed to do so.

While such machine-driven models had success in the past as **historical correlations** among different asset classes held firm, they have suffered in the wake of the pandemic as these linkages have **broken down**.

"Those who have instituted **careful data governance processes will be far more likely to succeed** in this game than those who haven't because rubbish in is rubbish out in the world of machine modelling," 日本ファイナンス学会第2回秋季研究大会

本日のお話

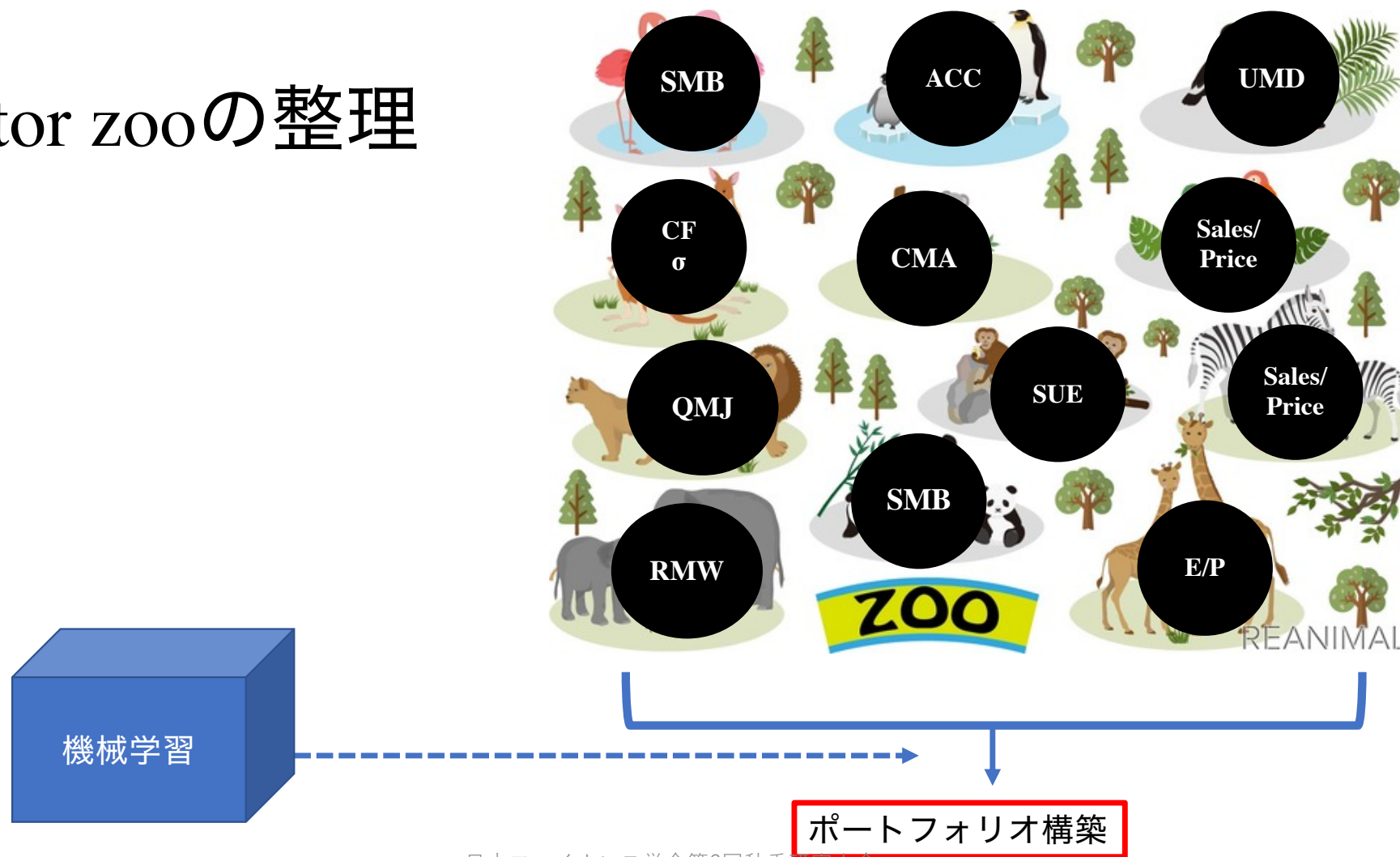
- i. オルタナティブ運用のために開発されているモデルの構築プロセスを中心に紹介
 - i. Long/Short Market Neutral
 - ii. Global Macro
- ii. 時間が許せば、tick dataとファクターを組み合わせた、マーケット状態把握のための機械学習モデル構想

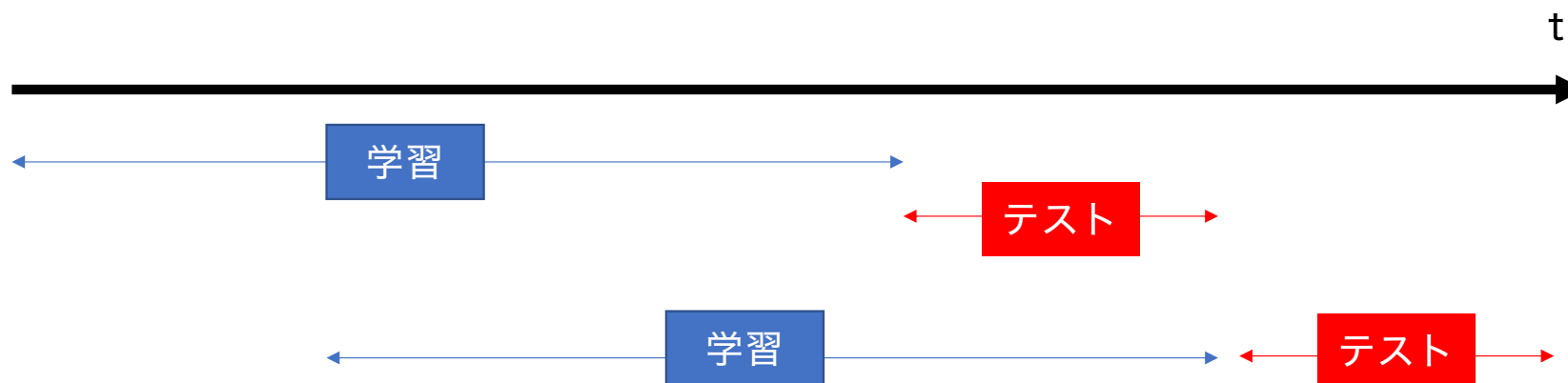
私達の立場

- 真に資産価格評価のメカニズムを明らかにすることよりも、予測を正しく行なうことに重きを置かざるをえない
- 資産運用のためモデル開発は”Data mining exercise”の側面が強くならざるを得ない

Long/Short Market Neutral Model

Factor zooの整理





今どのファクターに重みを乗せるべか？

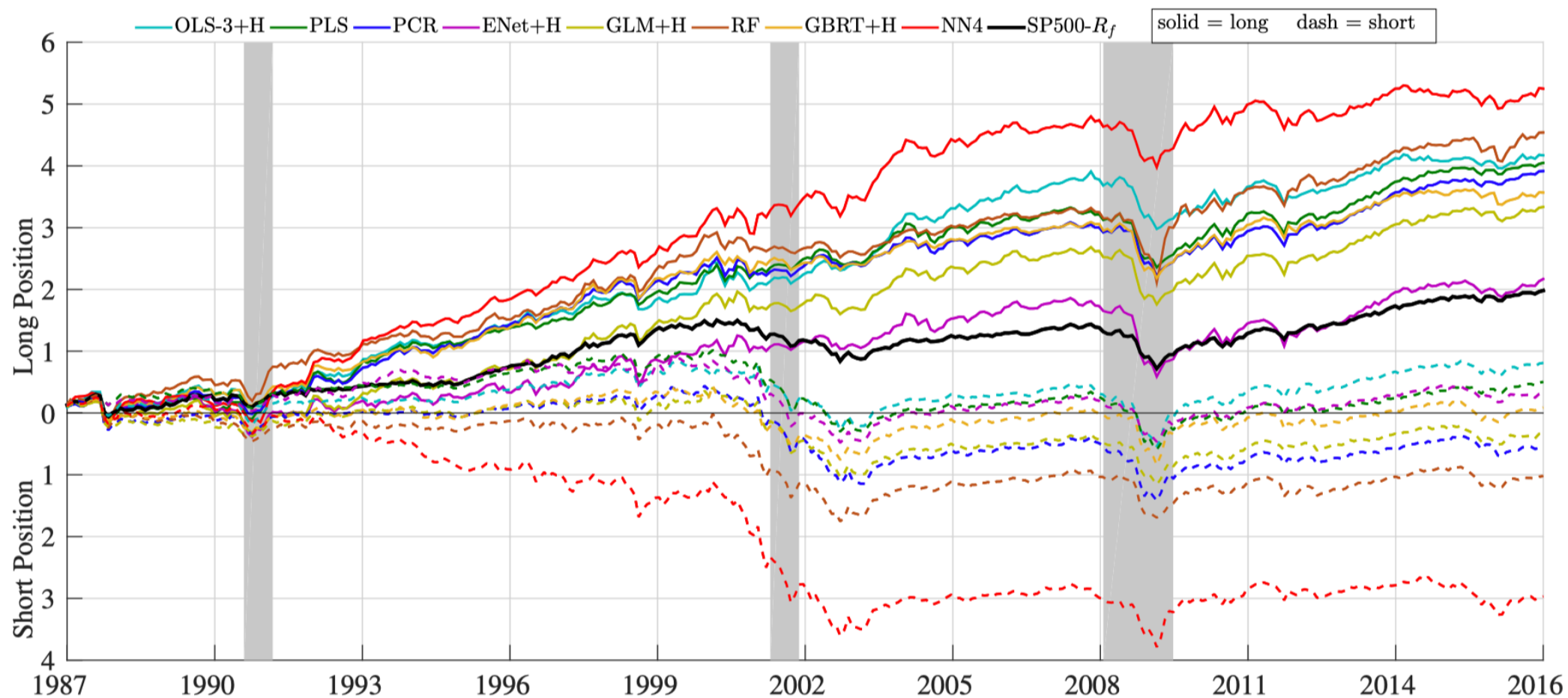
機械学習を活用することで予測力は向上するのか？

- Gu, Kelly, Xiu (2019), WP
OLS < LASSO, Elastic Net < Boosting, Neural Network
テストしたpredictor 900+
- Freyberger, Neuhier, Weber (2020), RFS
OLS < LASSO
テストしたpredictor 62
- 清水 /内山 (2020)
OLS (単純なz-score) = Boosting, Neural Network
テストしたpredictor: ValueとQuality

向上した

向上した

向上しない

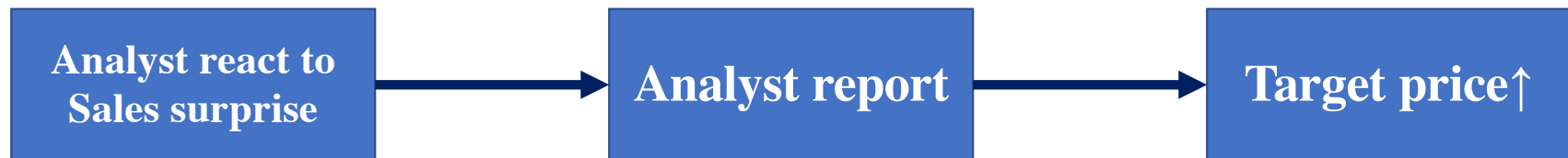


そもそも機械学習を実施する段階から任意の選択が始まっている

- どのような関数型をもちいるのか
- 学習セットとバリデーションセットを用意したとしても、うまくいかなければ、パラメーターチューニングをおこなう
- うまく機能するモデルは、結果的にはoverfitしている可能性が高い

- オルタナティブデータを用いて新たなファクターとして扱う

オルタナティブデータでオルタナティブファクターを作成する



- ✓ アナリストはとりわけ既存店売上（Same store sales, SSS）の変化率に対して業績予想を左右させている
- ✓ POS情報に基づいて銘柄をクロスセクションにランキングしL/Sのリターンを計測する。（これを*Alternative data based factor return (AFR)*と呼称する）

確認事項

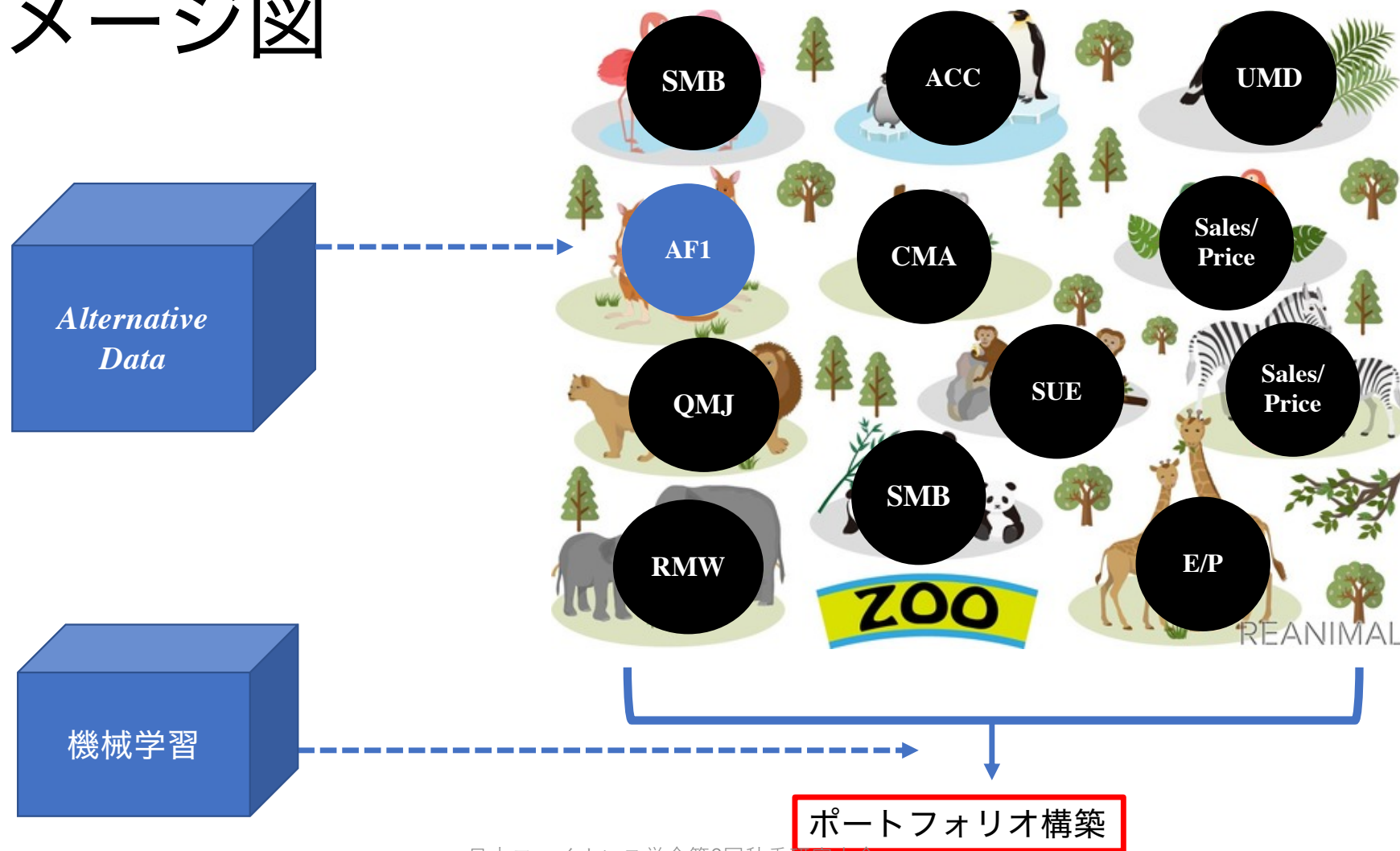
ある AFR が高いと確認できたとする。追加的に、以下の確認要

1. AFR は既知のファクターリターンと直交する性質を持つものなのか

例えばHeston and Sadka (2008)のSeasonality factorは3FFでは有意なファクターであるが、Carhartのモーメントとの相関が高い

2. AFR 評価の際、何をBenchmarkポートフォリオとするか考える。既存のスタンダードな資産価格評価モデルはomitted variable 問題に直面しているので使いにくい
3. 既知のファクター群から、学習によってBenchmark portfolioを作成し、 AFR に追加情報が含まれているかを確認する。

イメージ図

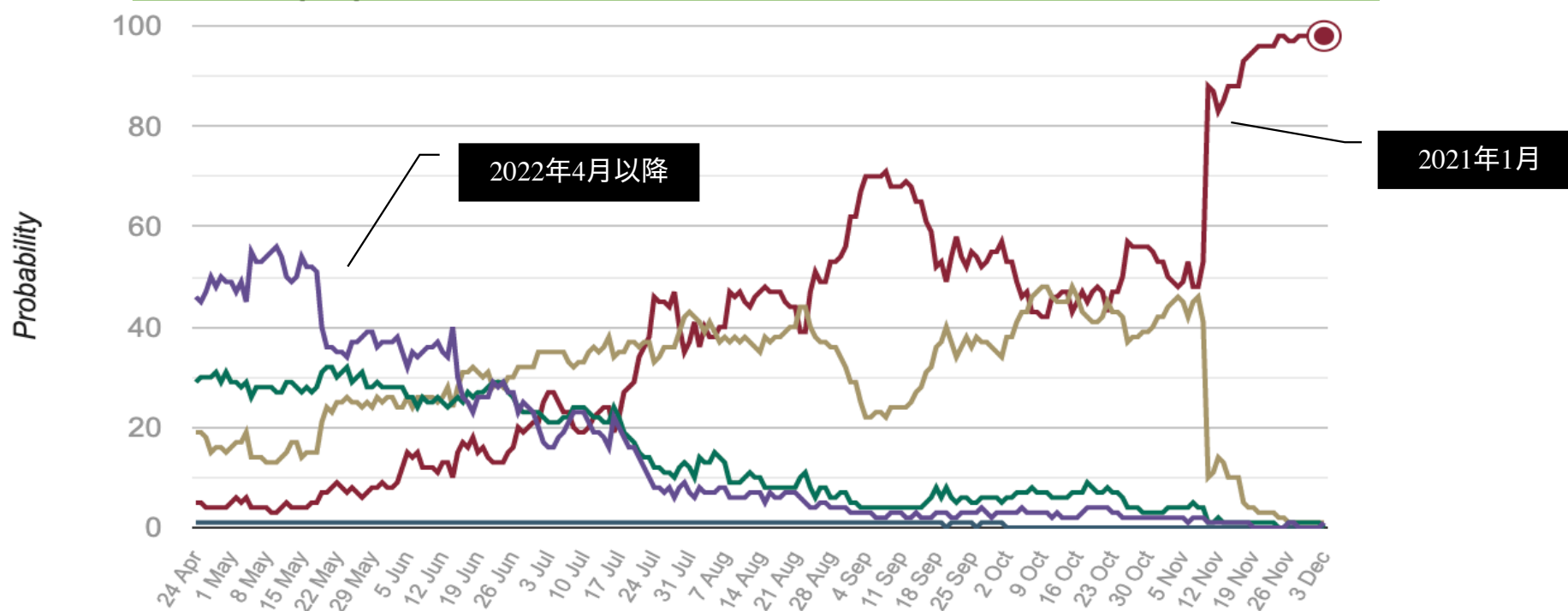


業種スペシフィックなオルタナティブデータの場合

- クロスセクションに十分なカバレッジがない場合、新たなファクターとして扱うことが困難
- 業種の重み調整として使う
 - 例) ゲーム

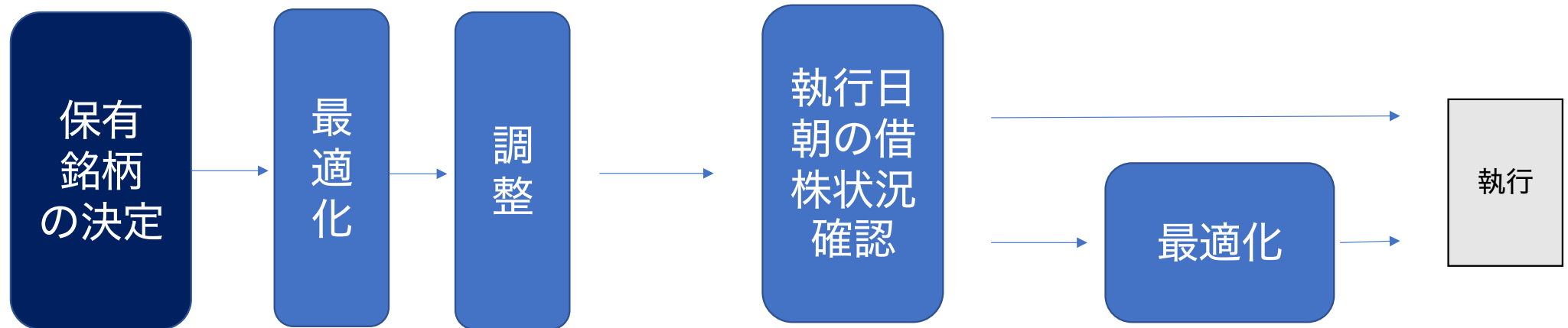
足下の状況において効果が高いオルタナティブデータ

米国において2500万人がワクチンを受けられるようになるのはいつか？



出所:Good Judgement Inc. Hpより

- 銘柄・業種、ファクターへのエクスポージャーの決定
- 取引コスト・借株コスト等の制約条件を付加して最適化



- オルタナティブデータによる調整

- 最適化後のポートフォリオで借株料の上限を満たしているか？
- 空売りしたい数量の確保は可能か？

- 問題ない場合は執行へ
- 借りられない等の問題があるとき Short側の制約を満たすように再度最適化

Global Macro Model

Cross sectional predictorの中でも、概ねどのモデルでも選ばれ、強く影響しているのは、価格トレンド、モーメンタムやリターンリバーサル、売買回転率である

テクニカルシグナルの生成

変数名	定義	ρ	観測日数(d)
$close_t$	c_t	0.8	10
$kerb_t$	o_t/c_{t-1}	0.9	9
$barb_t$	c_t/o_t	0.9	9
$range_t$	$h_t - l_t$	0.8	10
$volume_t$	v_t	0.9	10
kMb_t	$kerb_t - barb_t$	0.9	10
$shadowU_t$	$(h_t - \max(o_t, c_t)) / range_t$	0.8	10
$shadowD_t$	$(\min(o_t, c_t) - l_t) / range_t$	0.8	10
$shadow_t$	$shadowU_t * shadowD_t$	0.8	10
$aReturn_t$	$c_t/c_{t-1} - p_t/p_{t-1}$	0.8	9

日次データからその日の特徴量の抽出 - 1

■ここでは枝密度を特徴量とする

$$E_t = \{(a, b) | \text{sim}(\mathbf{x}_{t,d}^a, \mathbf{x}_{t,d}^b) \geq \rho; a, b \in N_t\}$$

過去d日間の時系列データにおいて、任意の2銘柄a,bの類似度を相関係数で測定する。2銘柄には任意の相関係数(ρ)以上が観察された場合にのみ枝がはられる。その日のマーケットの特徴量を枝密度として定義する

日次データからその日の特徴量の抽出 - 2

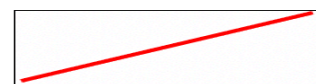
■どの局面の特徴量であるかを定義する

$$\text{sim}(\mathbf{x}_{t,d}^a, \mathbf{s}_{t,d}) \geq \sigma \quad \text{and} \quad \text{sim}(\mathbf{x}_{t,d}^b, \mathbf{s}_{t,d}) \geq \sigma$$

先の条件に加えて、個別銘柄の動向については追加的条件を付加する。銘柄が相場が上昇（下落）局面の時の類似度であるのか、急上昇、急下降時の類似度であるのかを判断させる。

具体的には、以下の局面ベクトルとの相関も条件に加えて、その日の枝密度を測定する

①上昇



up: lu

②下落



down; ld

③急上昇



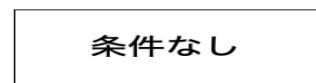
sudden - up; su

④急下落



sudden - down; sd

⑤条件なし（値動きの相関関係だけを見る）



条件なし

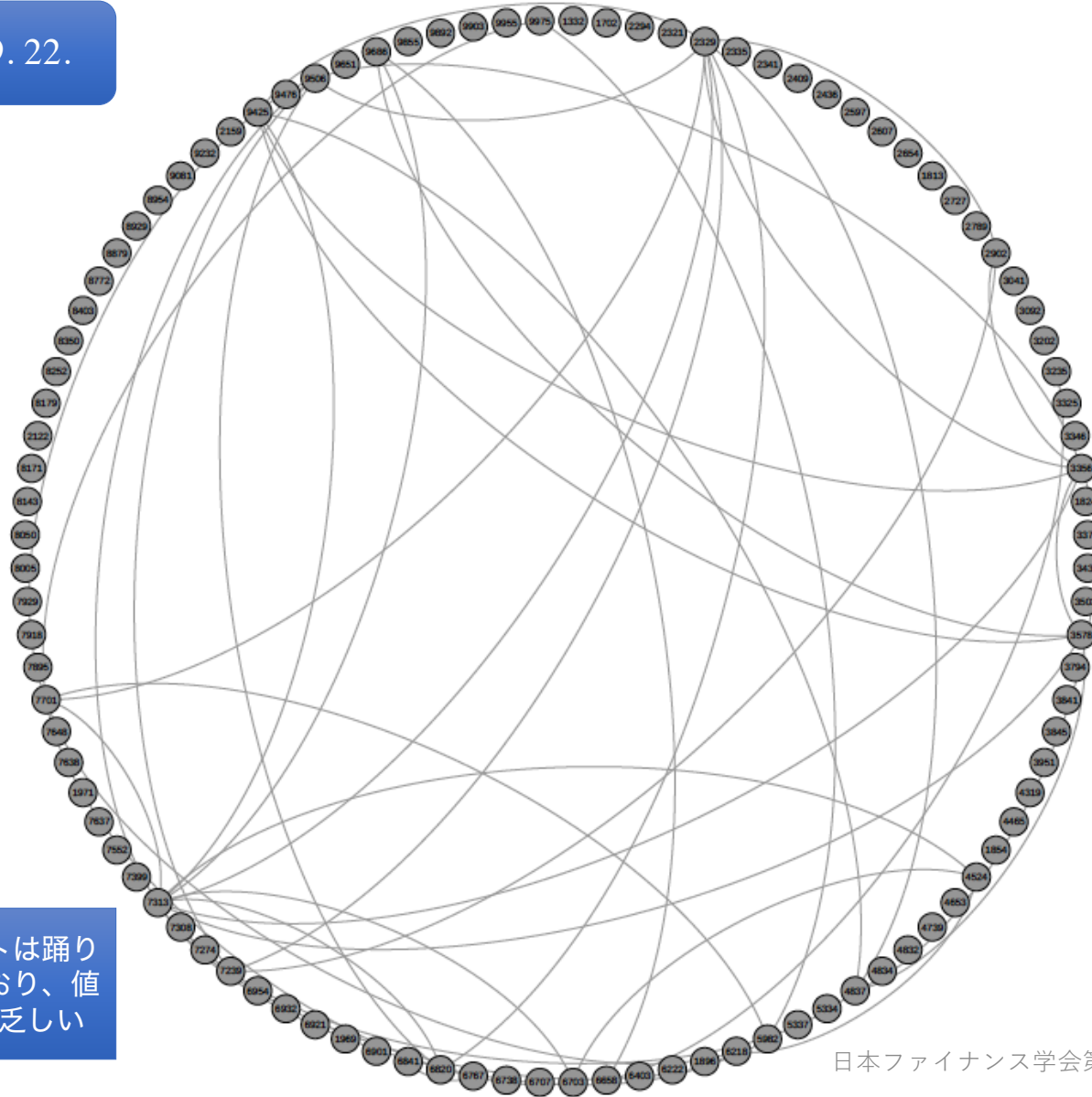
all

10種類×5局面=50 のシグナルの生成

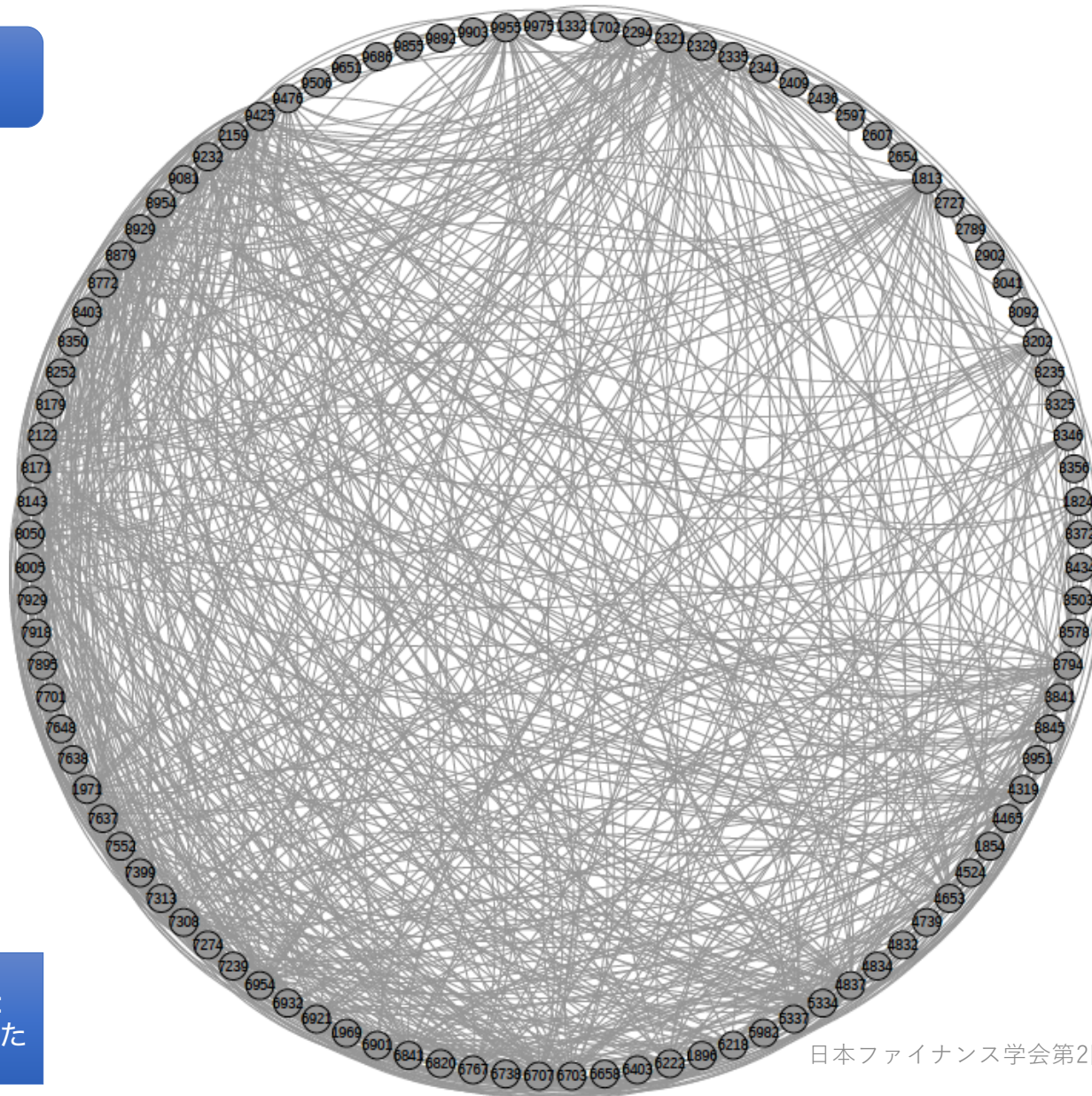
枝密度で投資タイミングを図るイメージ図

2008 9. 22.

マーケットは踊り
場にきており、値
動きには乏しい

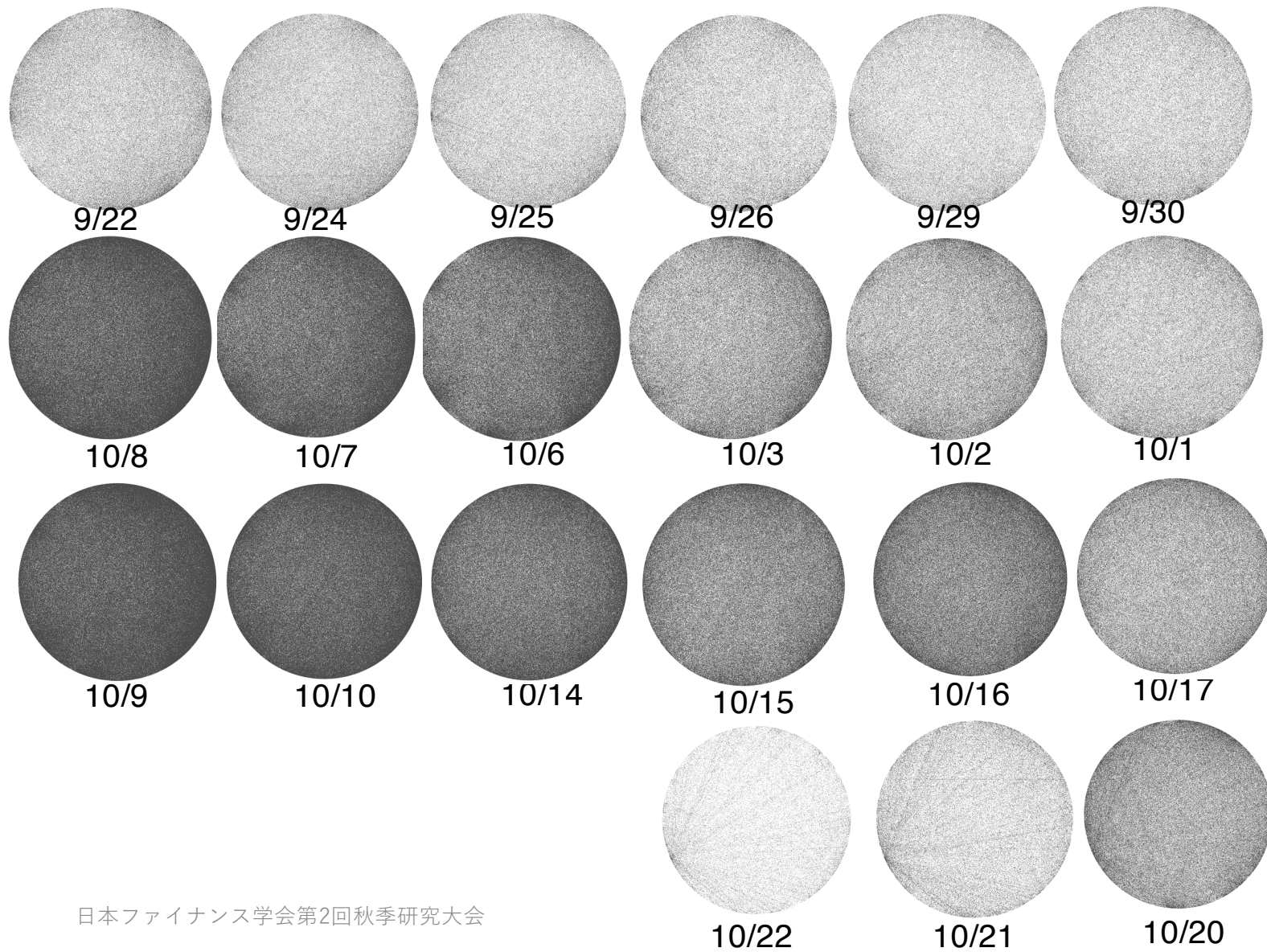


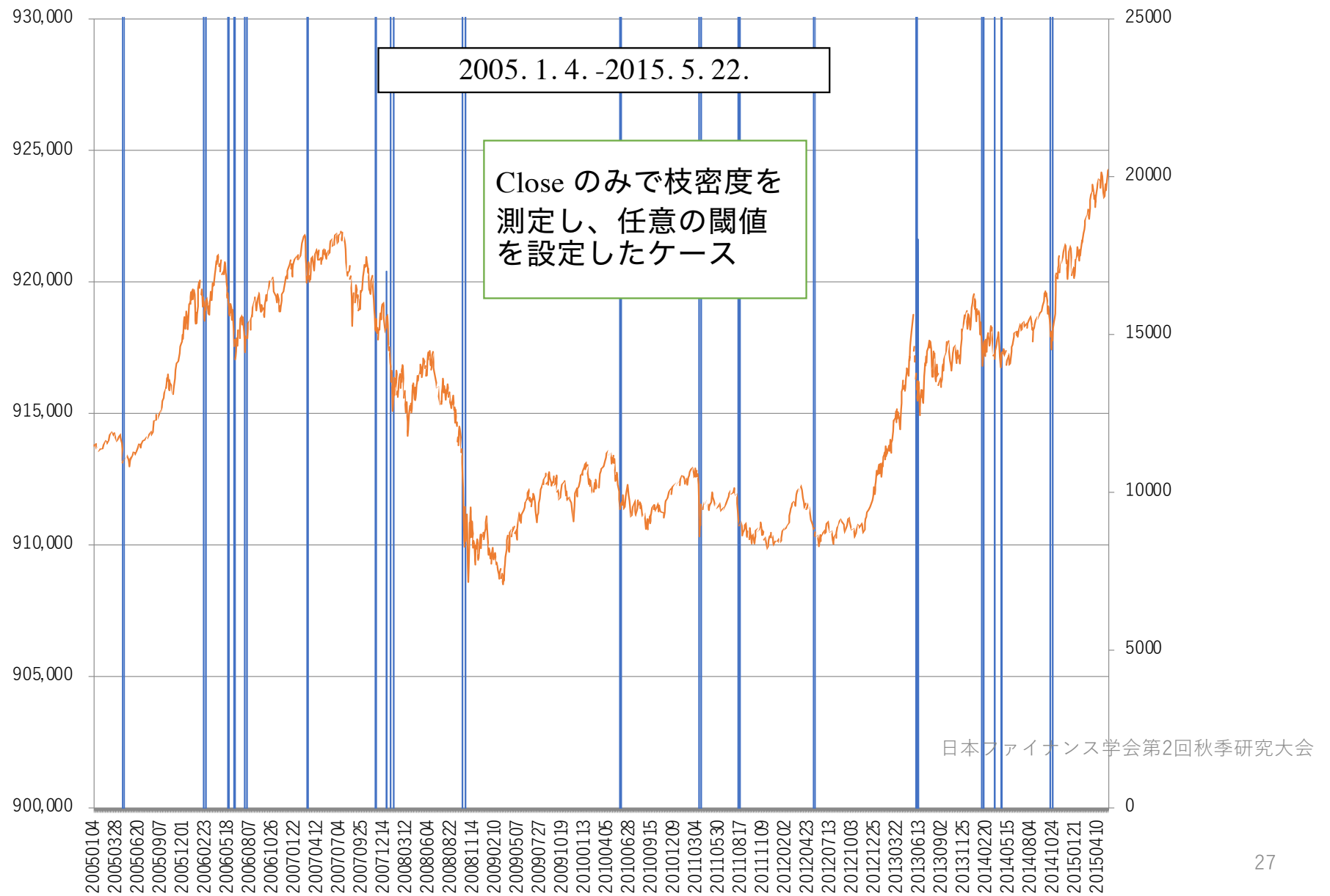
2008 10. 6.



マーケット状況:
大幅下落を記録した

Time series changes of complexion of stock universe surrounding Lehman debacle.





次の意思決定をデータマイニング的におこなう

- どのシグナルの枝密度を用いるべきか？
Ex: $close_t / kerb_t / barb_t \dots$
- どの局面の枝密度を用いるべきか？
Ex: $close_t - su / kerb_t - all / \dots$
- シグナルの組み合わせ効果は？
Ex: $kerb - all \& range - all$
- 枝密度の閾値設定基準は？
- Trading 戦略として実装する際に、何日positionを維持する？

単一のシグナル(single signal)の特徴だけを見る場合と、
2つのシグナル(double signal)の特徴を併せ見る場合を考える

$$Q_t(v) = (f_{t-1}^v < f_t^v) \wedge (f_t^v > f_{t+1}^v)$$

$$Q_t^m(v) = \bigvee_{i=t-m+1}^t Q_i(v)$$

- Single signal $SS_t(v) = Q_t^1(v) \wedge (f_t^v \geq \alpha_t^{s,k})$
- Double signal $DS_t^m(v_1, v_2) = SS_t(v_1) \wedge Q_t^m(v_2) \wedge (v_1 \neq v_2)$

Trading parameterの探索

■s: 過去どの程度の期間において枝密度の閾値を算出するか

$$s = 50, 100, 200 \text{ days}$$

■k: 期間sにおいて、上位k番目の閾値を使用する

$$k = 1, 5, 10, 20 \text{ th density}$$

■h: どの程度の期間保有するのか

$$h = 1, 2, 3, 4, 5 \text{ days}$$

■シグナル数

$$50 \times s(= 3) \times k(= 4) \times h(= 5) = 3,000$$

$$[50 \text{ (original signal)} + 2450 \text{ (Combination of } 50 \times 50)] \times s(= 3) \times k(= 4) \times h(= 5) = 150,000$$

2010-16年全期間
各国シングルシグナルの上位 (p-value <0.00001)列挙

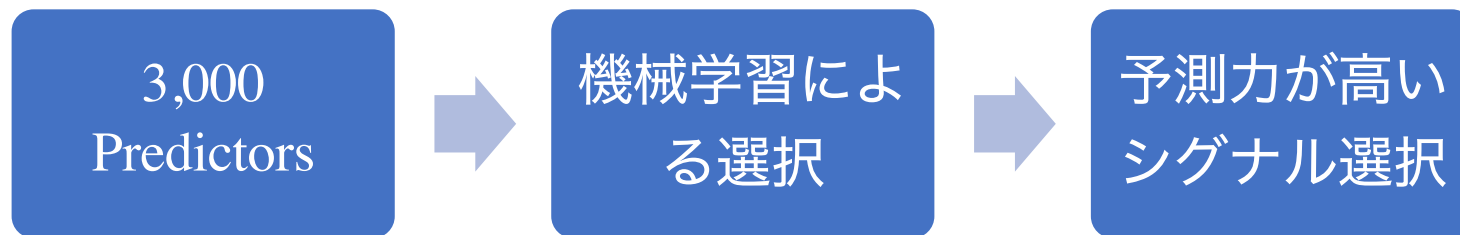
国	シグナル	h	s	k	件数	return(%)
jp	barb-all	2	50	5	124	0.88
	barb-sd	3	200	10	55	1.68
us	shadowH-lu	3	100	20	216	0.52
	shadowL-ld	5	50	20	452	0.38
	kMb-sd	5	50	1	53	1.55
gb	barb-su	1	200	15	60	0.61
	kMb-sd	2	200	20	90	0.75
	kerb-ld	4	50	15	185	0.55

(出所) 岡田・羽室 (2017)

2010-16年全期間 各国ダブルシグナルの上位 (p-value <0.00001)列挙

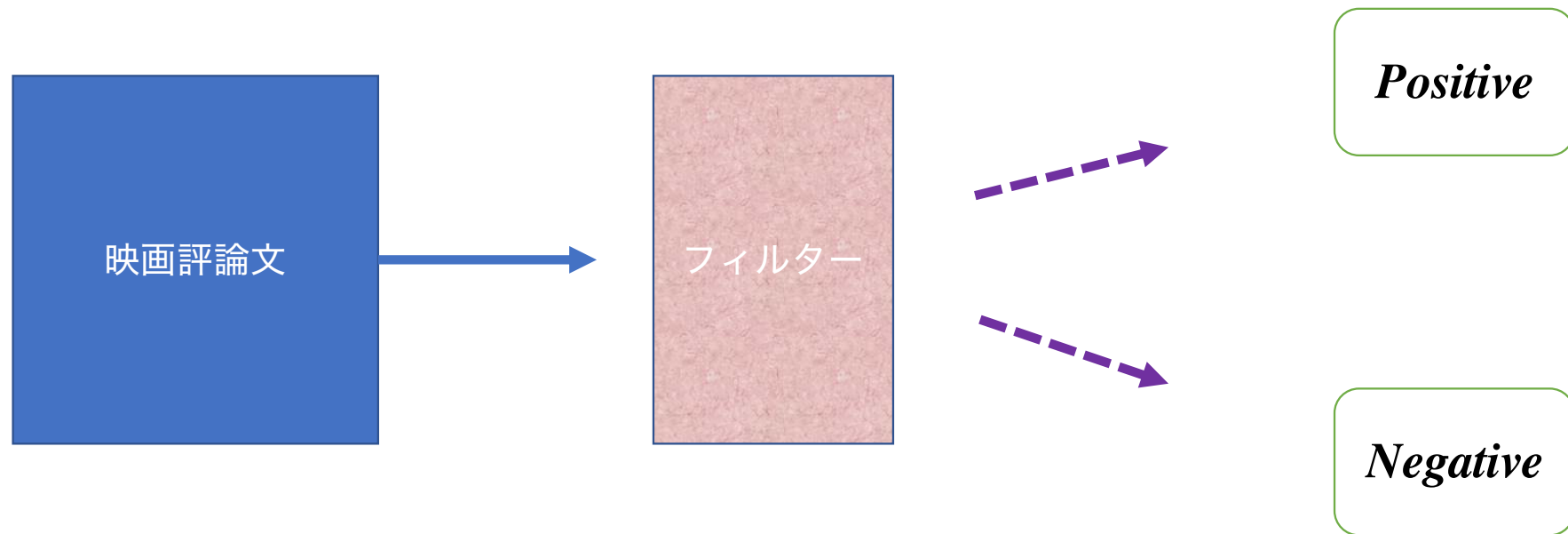
国	mainSignal	subSignal	hp	<i>s</i>	<i>k</i>	件数	return(%)
jp	kerb-all	range-all	3	50	15	50	1.24
	shadowH-su	kMb-all	1	50	15	91	0.52
	close-su	shadowH-lu	2	50	15	73	0.50
	shadowH-su	shadow-su	1	50	15	113	0.48
	shadowH-su	kMb-sd	1	50	15	102	0.48
	shadowH-su	aReturn-lu	1	50	15	101	0.42
	kerb-su	range-ld	2	50	20	62	-0.71
	kerb-sd	barb-sd	2	50	15	61	-0.86
	shadowH-ld	range-lu	5	50	20	65	-0.97
us	kerb-all	kerb-su	5	50	15	71	0.98
	kerb-all	kMb-lu	4	50	15	59	0.95
	kerb-all	range-lu	4	50	15	53	0.90
	kerb-all	volume-su	4	50	15	62	0.87
	kerb-all	kerb-ld	4	50	15	56	0.80
gb	shadow-su	range-su	5	100	10	51	1.12
	shadowH-lu	aReturn-sd	5	100	20	76	0.97
	shadowH-lu	shadow-ld	4	50	20	71	0.84
	kerb-sd	shadow-su	3	50	20	76	0.60

Single signal selection



シグナルの意味を考えるためのプロジェクト

映画評論文のpositive/negative判断



顕在パターン

表 1: $\sigma = 200, \gamma = 4$ の条件における顕在パターン上位 5 つ. 1 単語のパターン (上段) と 2 単語のパターン (下段).

ポジパターン			ネガパターン		
パターン α	<i>sup</i>	<i>GR</i>	パターン α	<i>sup</i>	<i>GR</i>
delightful	2.1%	6.8	waste	10.2%	10.4
beautifully	2.8%	5.9	pointless	3.6%	10.4
touching	2.8%	5.3	worst	16.3%	9.0
superb	5.0%	5.2	poorly	4.5%	8.3
magnificent	1.8%	4.8	awful	11.0%	8.0
highly recommended	1.8%	22.2	crap worst	1.8%	76.3
great superb	2.0%	8.4	waste worst	2.0%	63.3
excellent young	1.8%	8.4	awful worst	2.5%	31.8
excellent performances	1.7%	8.0	stupid worst	2.1%	29.0
love wonderful	3.4%	7.6	worse worst	2.2%	24.7

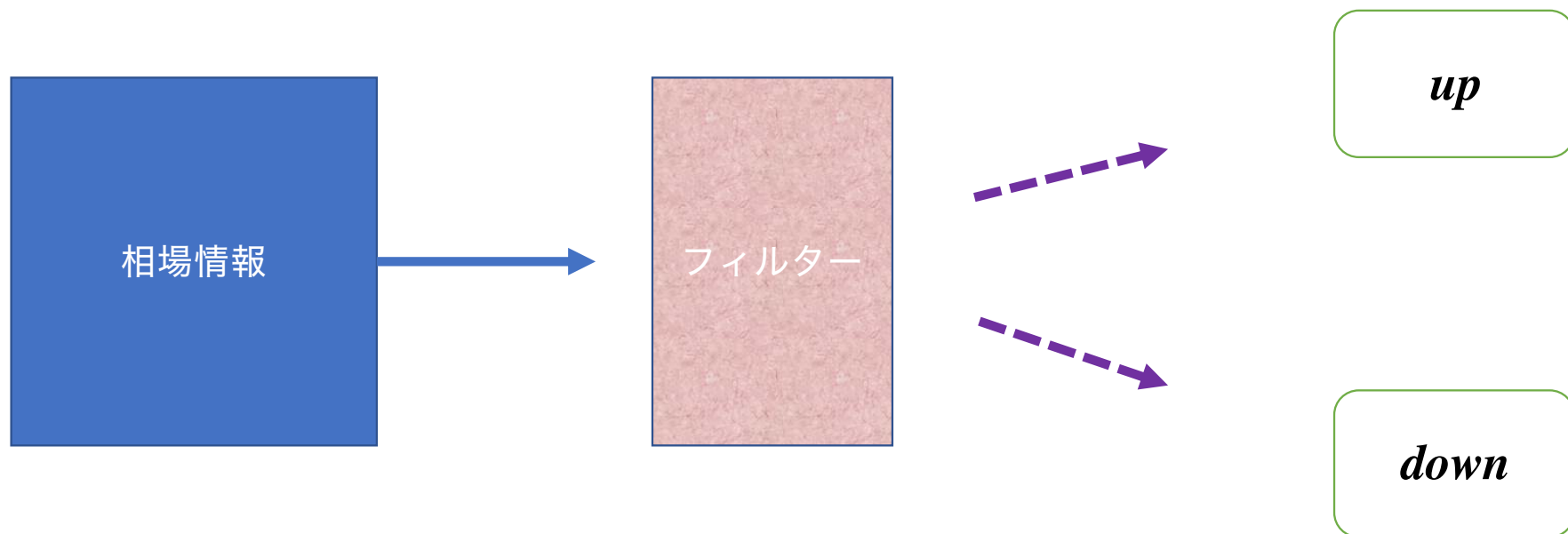
(出所) 羽室・岡田・寺田 (2021)

文章のpositive / negative判断を行なうが、どの部分にattentionしたかがわかる

previous reviewer claudio car gave a much better rec of the film 's plot details than i could . what i recall mostly is that it was just so beautiful , in every sense - emotionally , visually , editorially - just gorgeous . if you like movies that are wonderful to look at , and also have emotional content to which that beauty is relevant , i think you will be glad to have seen this extraordinary and unusual work of art . on a scale of 1 to 10 , i 'd give it about an 8 . 75 . the only reason i shy away from 9 is that it is a mood piece . if you are in the mood for a really artistic , very romantic film , then it 's a 10 . i definitely think it 's a must - see , but none of us can be in that mood all the time , so , overall , 8 . 75 .

it may treat important issues , yet not as a serious philosophy . it 's really difficult to care about the characters here as they are not simply foolish , just missing a spark of life . their actions and reactions are wooden and predictable , often painful to watch . the makers of earth know it 's rubbish as they have to always say " gene rod den berry 's earth . . . " otherwise people would not continue watching . rod den berry 's ashes must be turning in their orbit as this dull , cheap , poorly edited (watching it without ad vert breaks really brings this home) tr ud ging tr aba nt of a show lumber s into space . spoil er . so , kill off a main character . and then bring him back as another actor . je ee z ! dallas all over again .

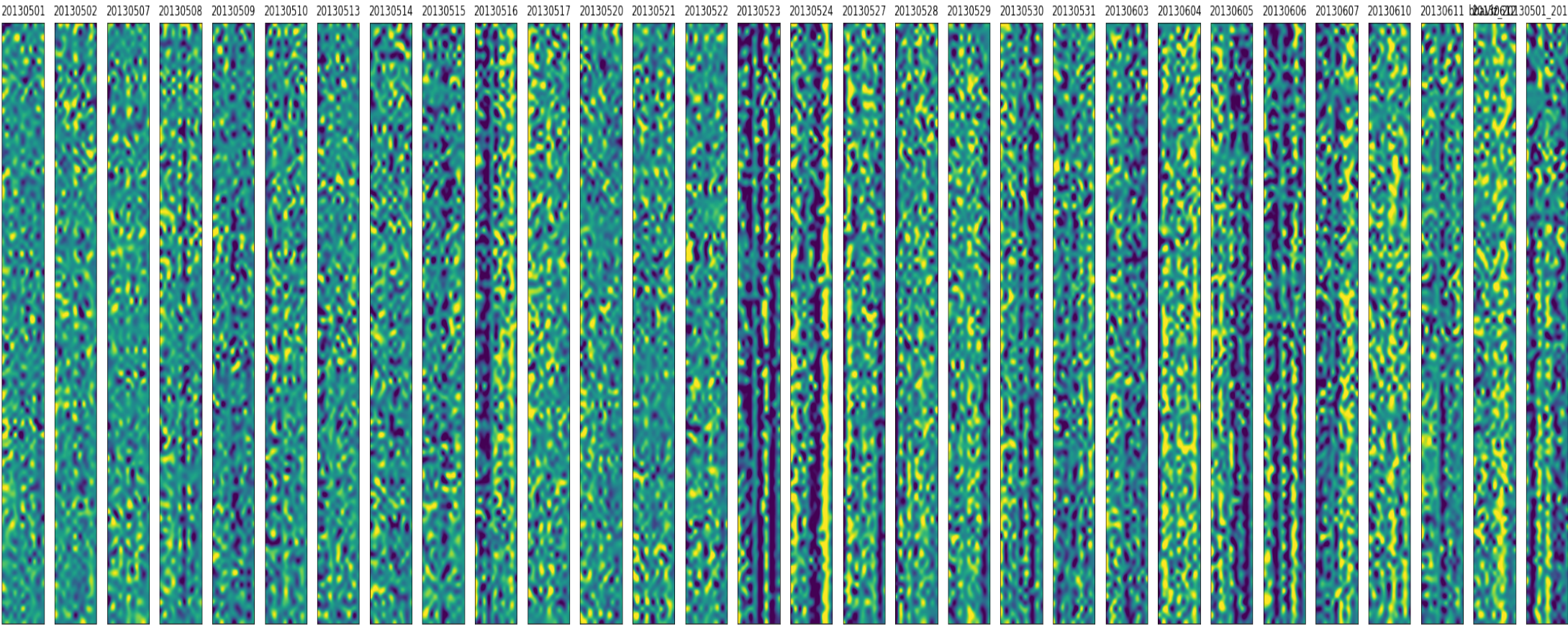
相場情報入力後の翌日の上昇/下落 判断



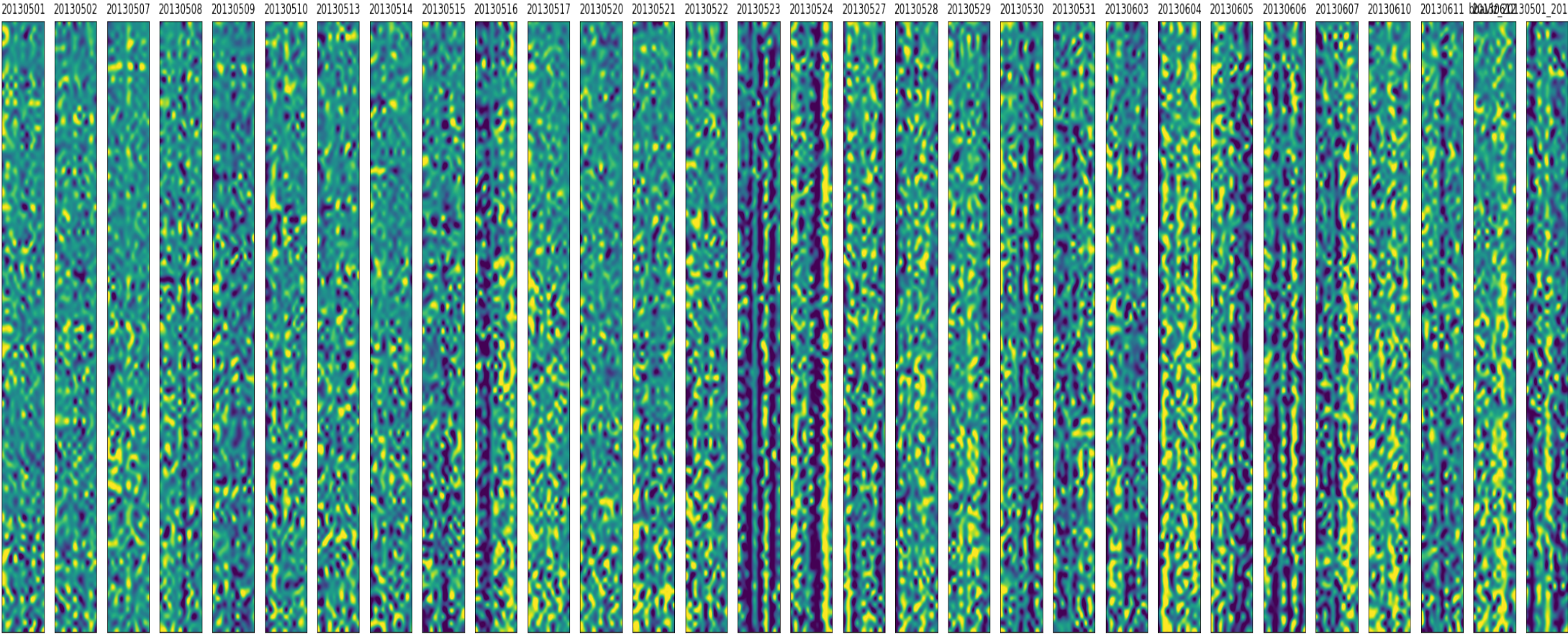
相場の上下判断の学習を考える

- 言葉に対してtick data (30分足 11次元/一日) を入力
- Size tick data は縦に規模順に並ぶリターンデータ
- 同様に、合計6ファクターについて実験中

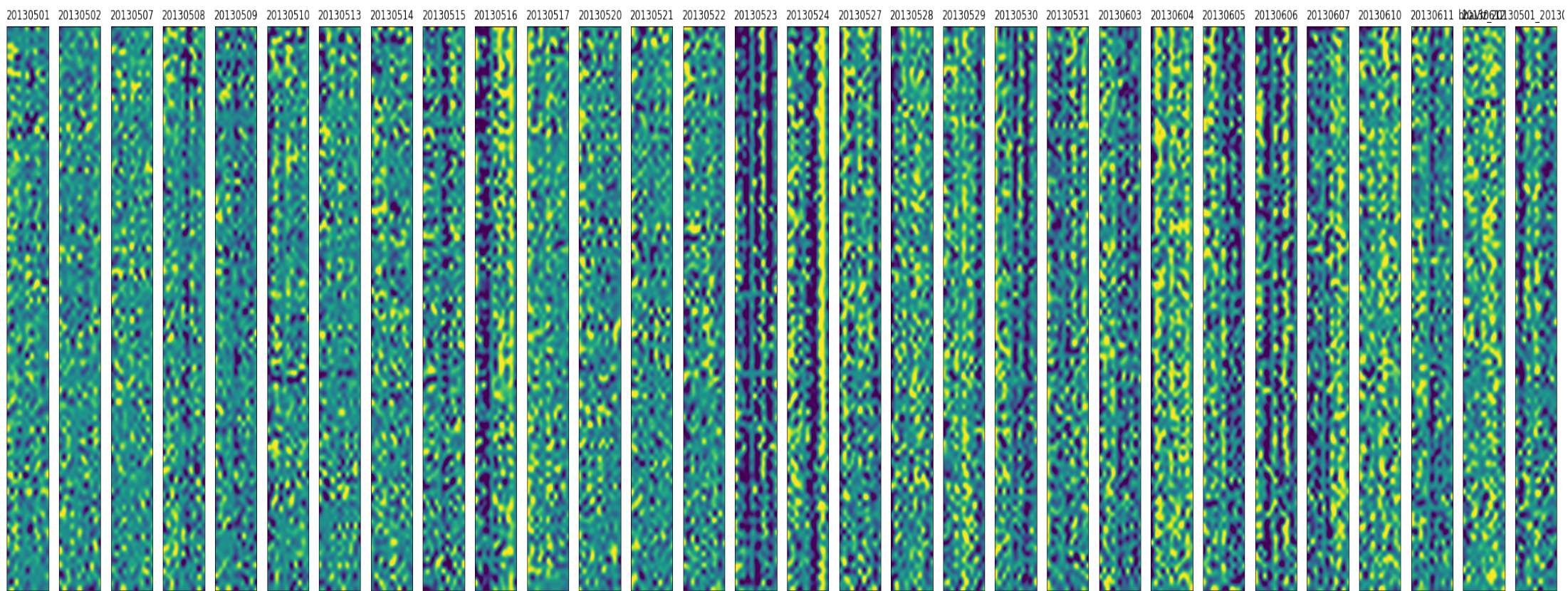
Size



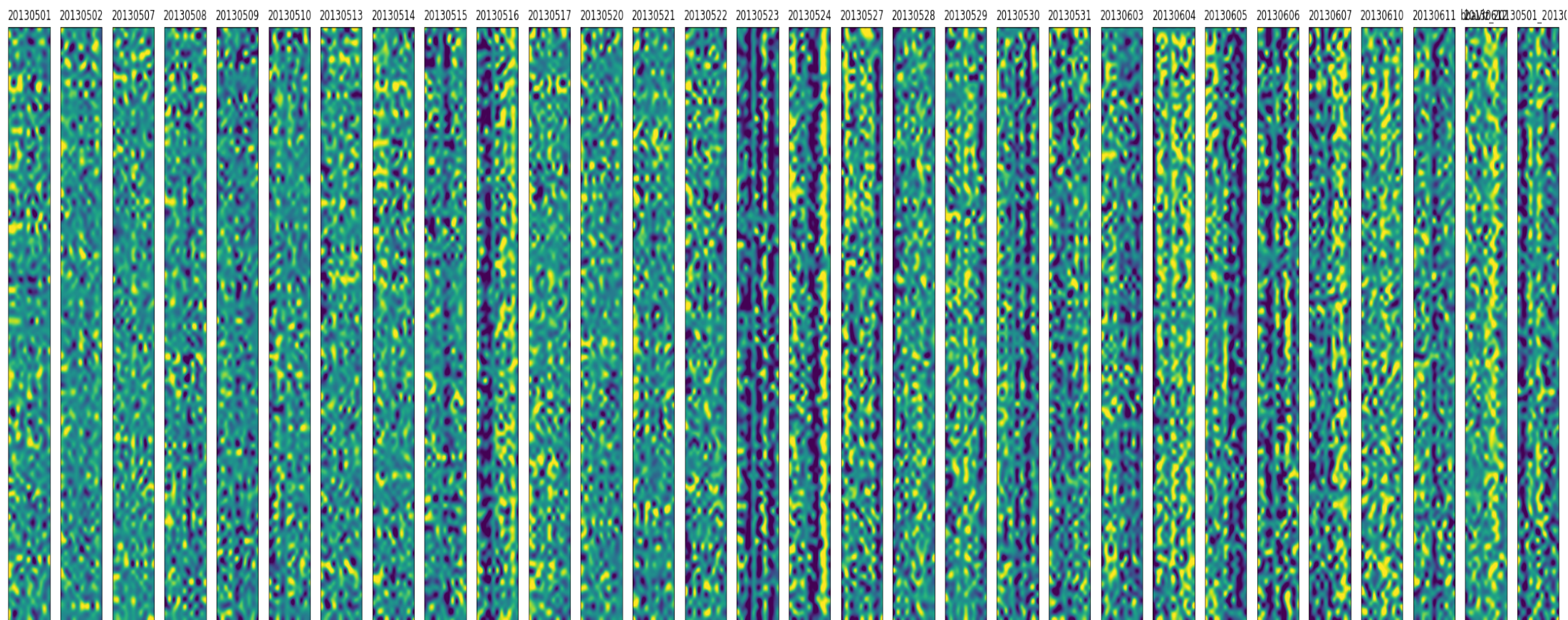
Beta



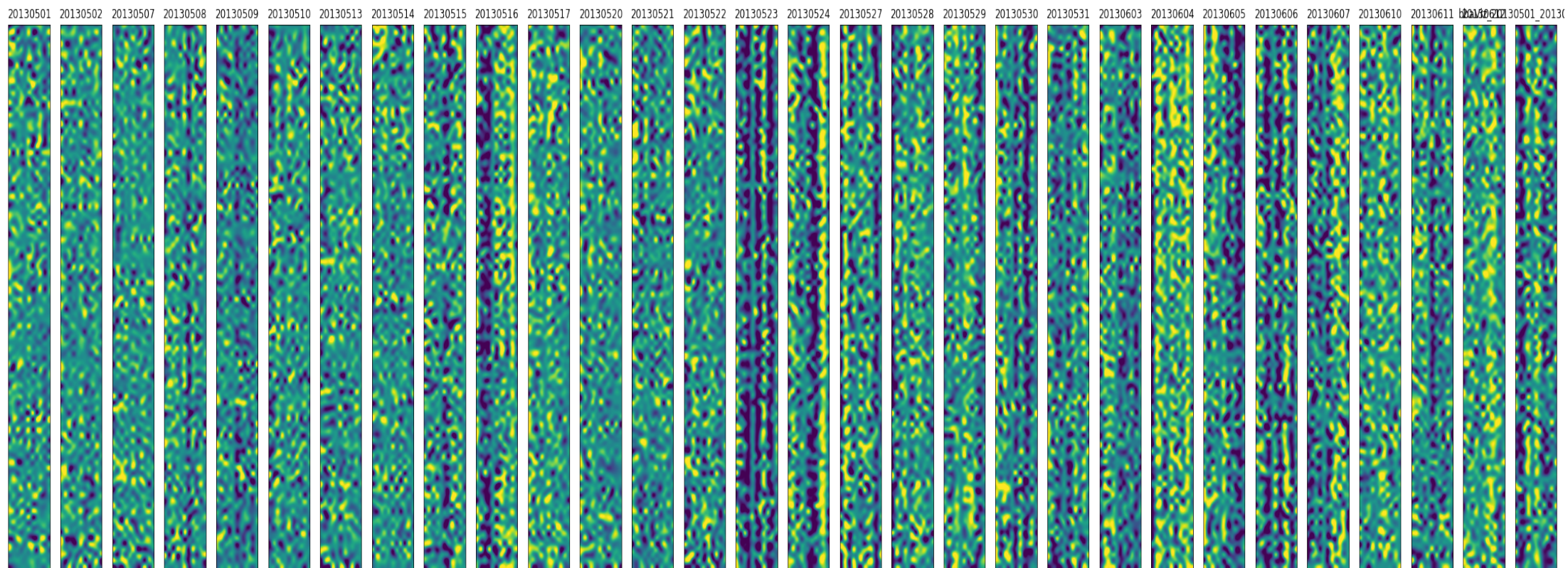
B/P



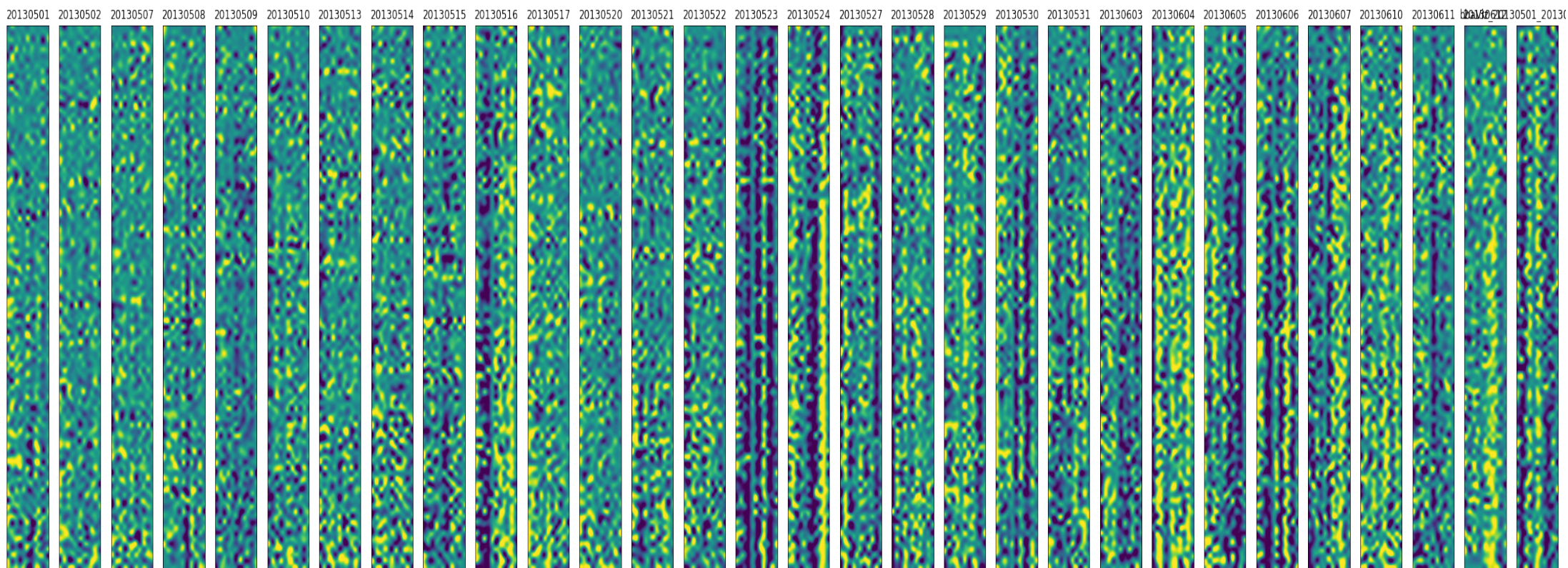
E/P



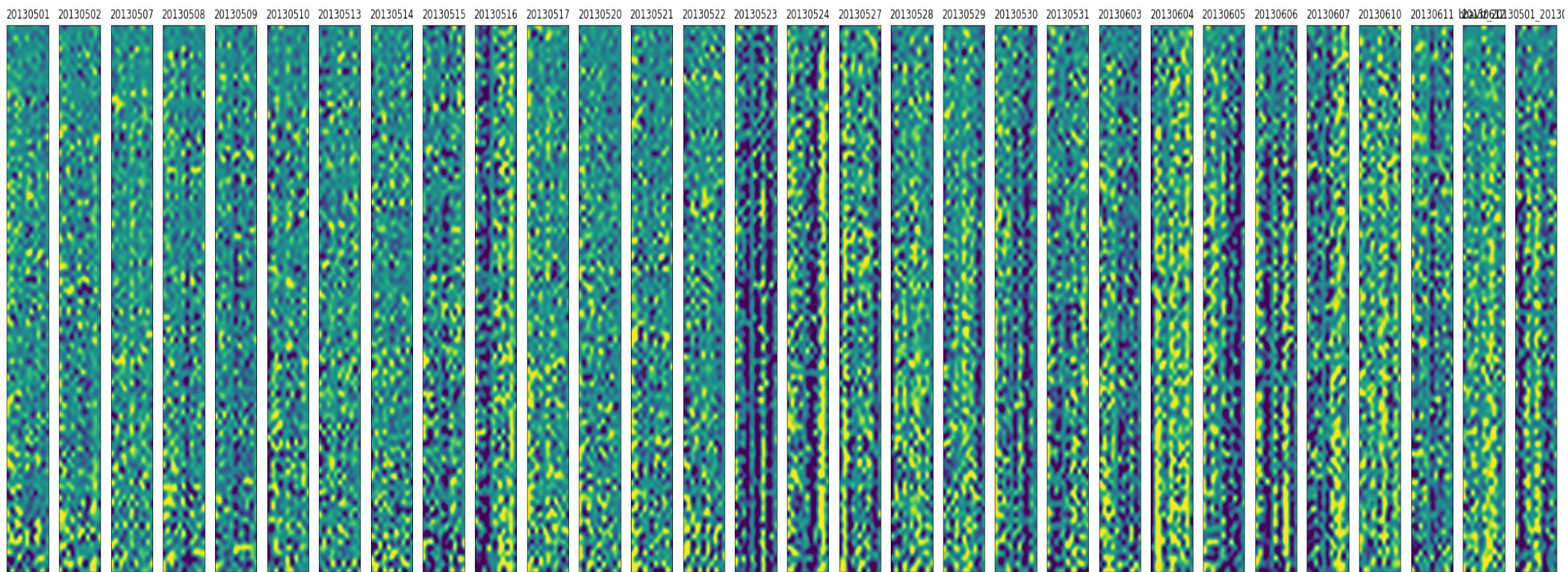
Sensitivity to US market



Turnover



i-vol



ご清聴ありがとうございました

参考文献

- 岡田・羽室(2017),「株式市場における大崩落兆し検知への挑戦」,第31回人工知能学会
- 羽室・岡田・寺田(2021),「社会科学におけるデータマイニングの役割」電気通信学会誌 2021 3月 *forthcoming*
- 清水・内山 (2020),「機械学習はファクターの合成方法を学習できるか」 2020年日本ファイナンス学会大会論文
- Feng, G, Giglio,S., and Xiu, D. (2019), Taming the Factor Zoo, Chicago Booth Paper No. 17-04
- Freyberger, J. , Neuhier, A., and Weber, M.(2020), ”Dissecting Characteristics Nonparametrically,” Review of Financial Studies, No.33.
- Gu, S., Kelly B. and Xiu D. (2018), “Empirical Asset Pricing via Machine Learning”, NBER Working Paper 25398