BIG DATAとAIで
マーケットの何を探すべきか？

関西学院大学大学院経営戦略研究科
Magne-Max Capital Management
岡田克彦

第４回金融資本市場のあり方に関する産官学フォーラム
@東京大学国際学術総合研究棟

既存のAIファンドは、AIをどう活用しているのか？

- 膨大なニュースの中から、関連性の高いニュースを抽出
  =ニュースを読まずに自動抽出
- リサーチリポート
  =中身を読まずに、センチメントスコアリング
- ディープラーニングによるスコアリング（？）
- ウェブデータ
  ＝注目度の代理変数
- ビッグデータに基づいた、短期予測、中期予測
- ビッグデータに基づいた、相場の転換点予測

どういう問題が起きているのか？

問題設定

説明変数と目的変数の間に
そもそも普遍的関係が
存在するのか？

問題の解き方

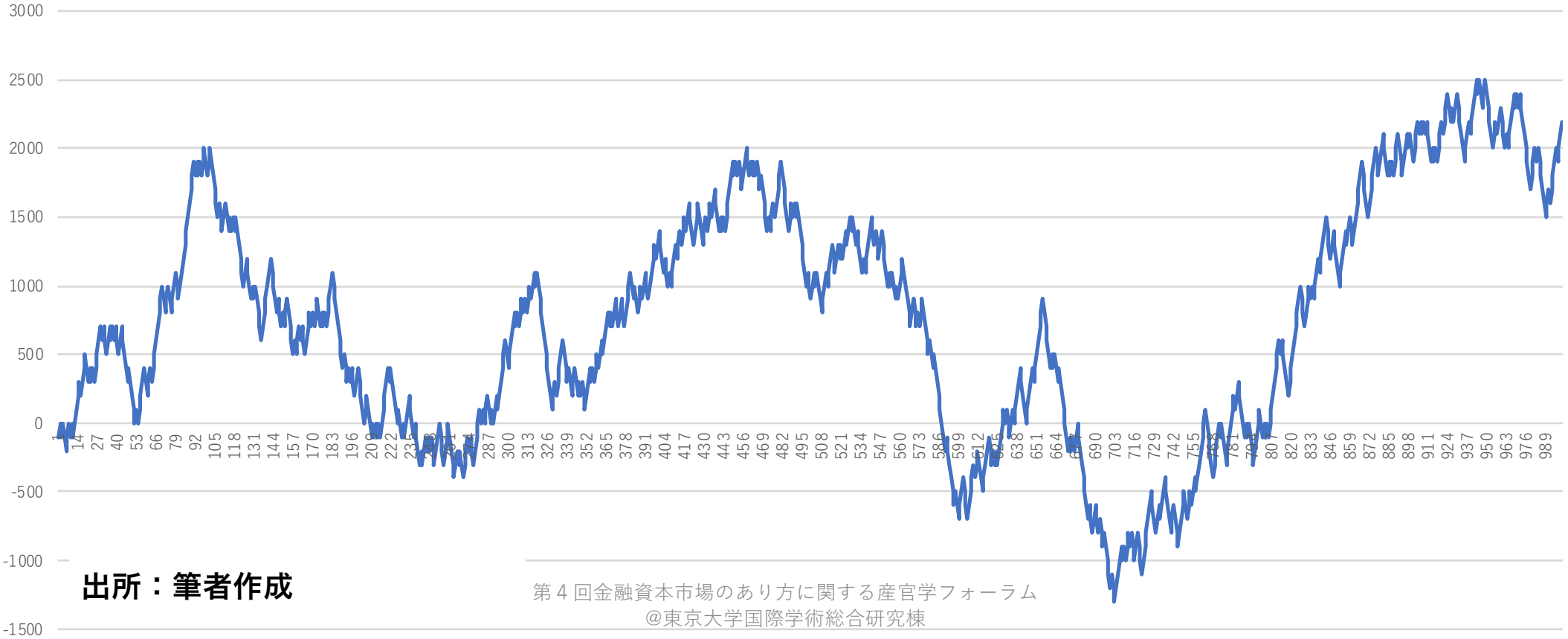
データ・ドリブンで関係性を学
習する際、過学習に陥って
いないか？

KendallやMalkielの言うように、株価系列がランダムウォークであるとしたら、それはAIが解く問題設定として適切か？

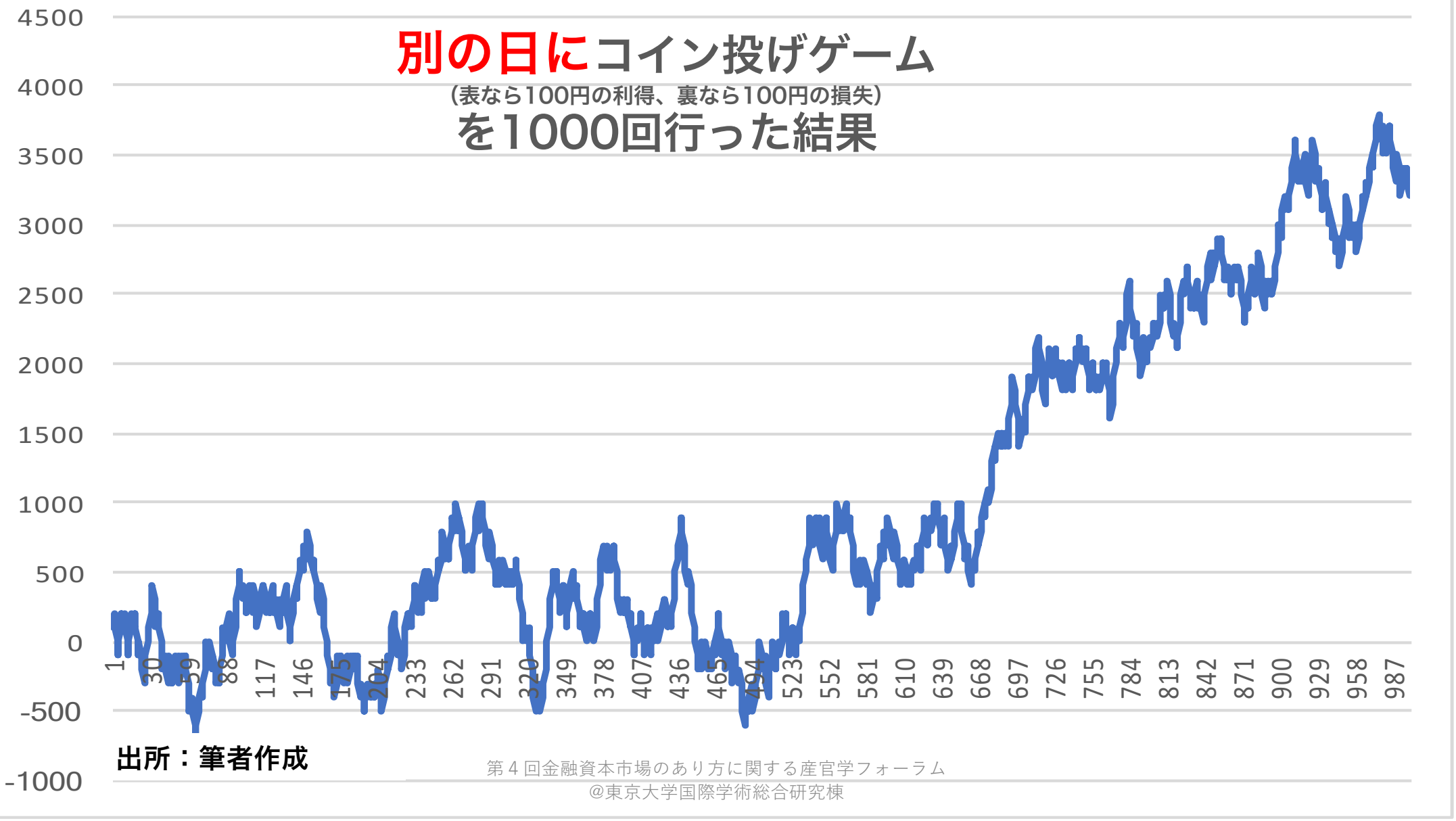
ある日コイン投げゲーム

（表なら100円の利得、裏なら100円の損失）

を1000回行った結果



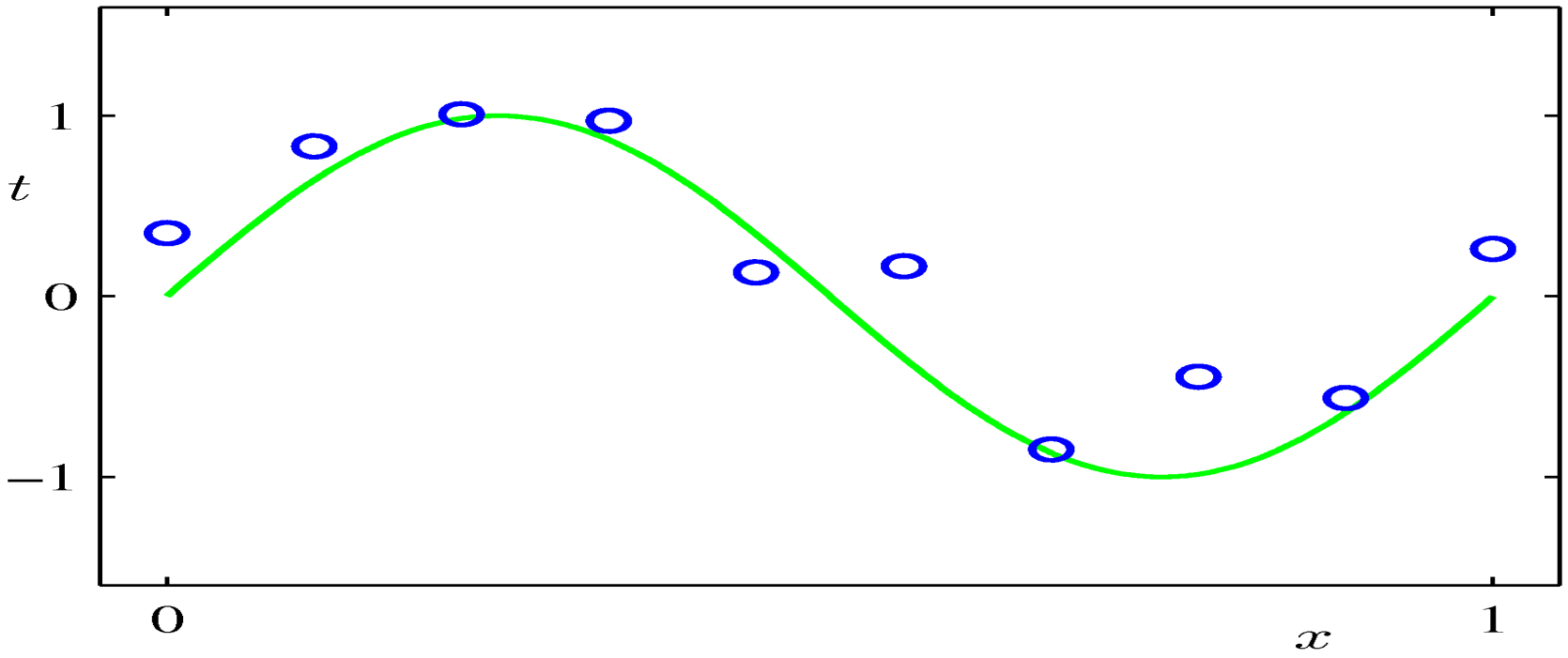
出所：筆者作成


別の日にコイン投げゲーム
(表なら100円の利得、裏なら100円の損失)
を1000回行った結果
4500
4000
3500
3000
2500
2000
1500
1000
500
0
-500
-1000
1
30
59
88
117
146
175
204
233
262
291
320
349
378
407
436
465
494
523
552
581
610
639
668
697
726
755
784
813
842
871
900
929
958
987

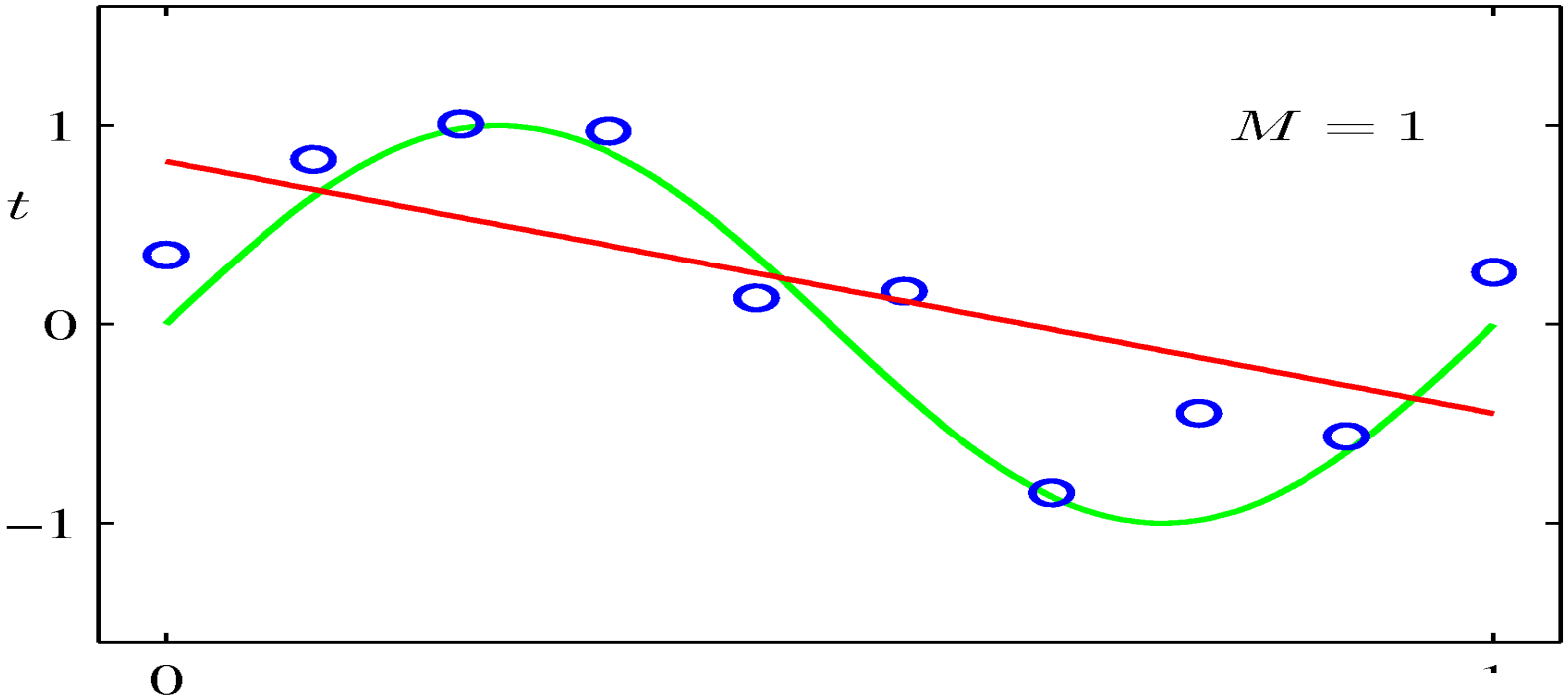
出所：筆者作成

モデルの汎化誤差は小さいものと言えるか

神のみぞ知る「特徴量」とマーケットの関数を推定したい



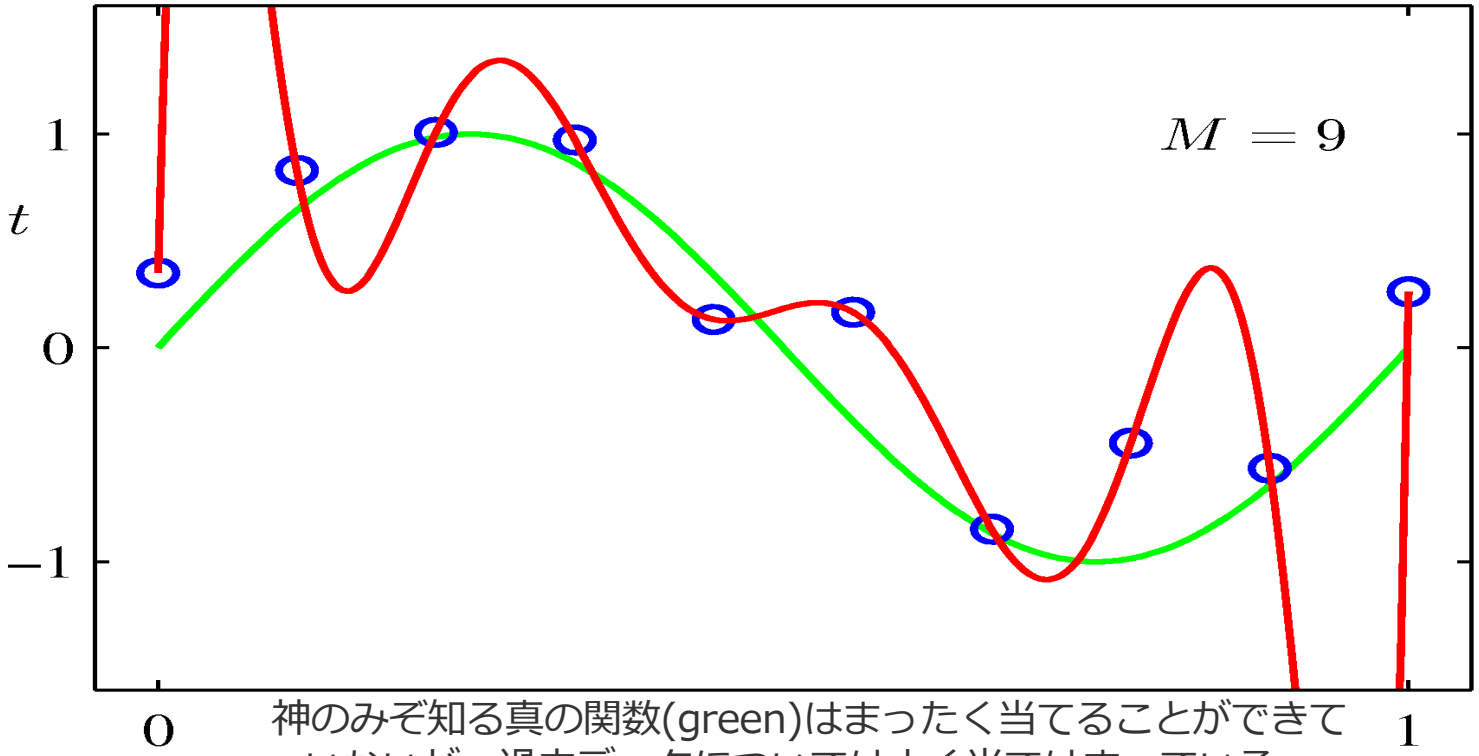
でも、線形回帰モデルだとこうなっちゃう



神様しか知らない関数 (green)をデータから推定しようとした。

でも、このモデルは “under fit” データとの誤差大きすぎ！

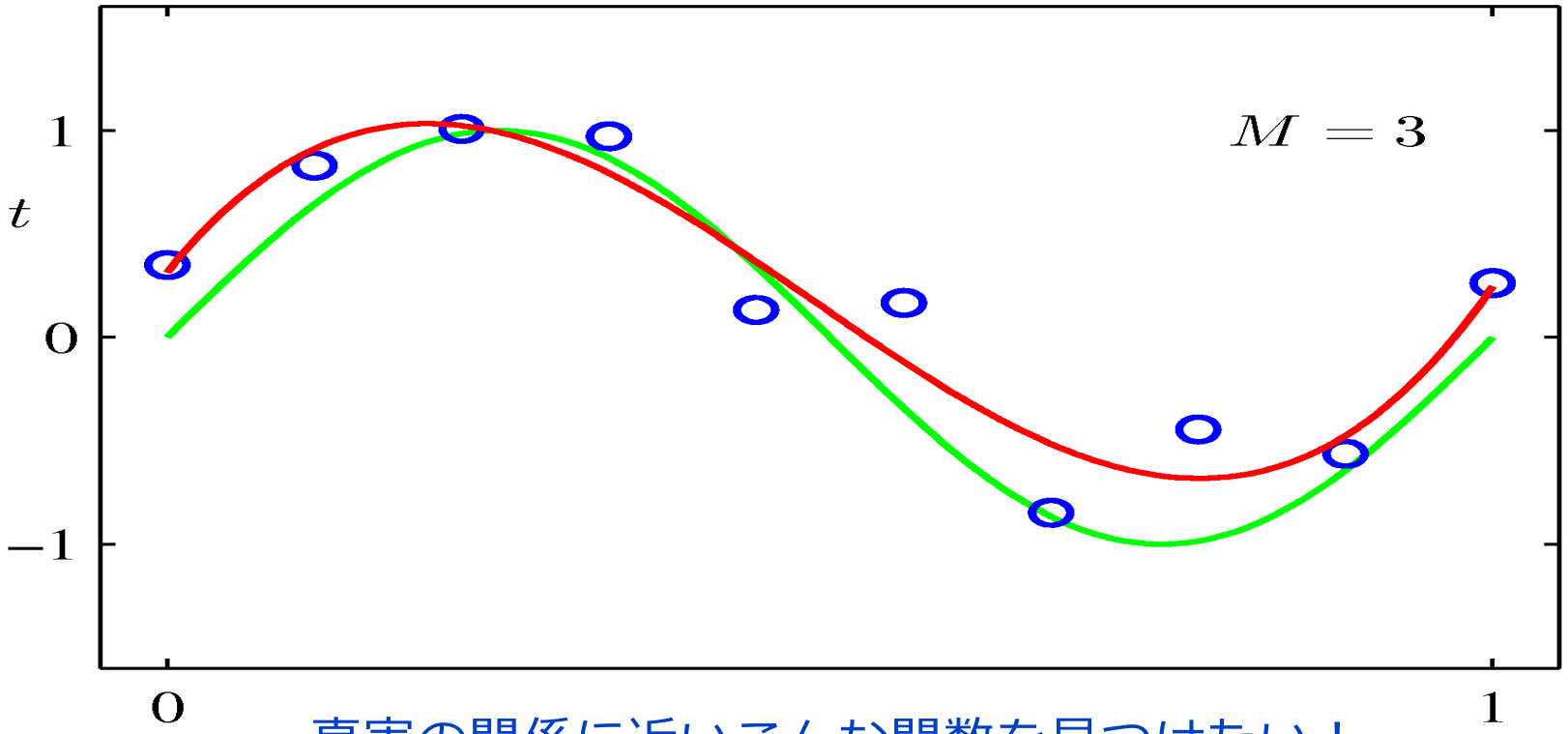
9次元関数にしたら、めちゃくちゃよく当たる！



神のみぞ知る真の関数(green)はまったく当てることができていないが、過去データについてはよく当てはまっている。

The model is “over fit”

3次元関数で推定してみたら・・・・

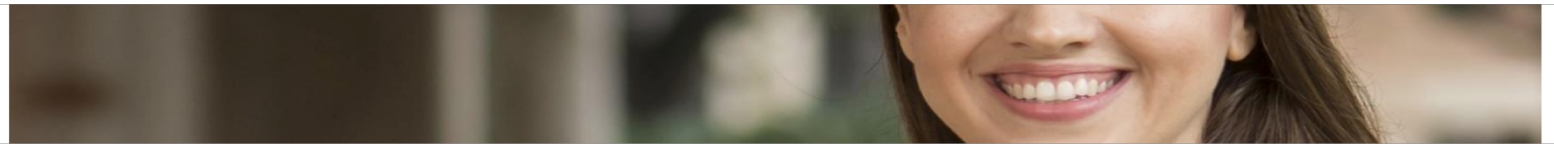


真実の関係に近いこんな関数を見つけたい！

Machine learning (ML) is a branch of statistics and computer science concerned with building computational systems that learn from data rather than following explicit instructions. Allen said much attention in the ML field has focused on developing predictive models that allow ML to make predictions about future data based on its understanding of data it has studied.

"A lot of these techniques are designed to always make a prediction," she said. "They never come back with 'I don't know,' or 'I didn't discover anything,' because they aren't made to."

She said uncorroborated data-driven discoveries from recently published ML studies of cancer data are a good example.



明確にすべき問題意識と方法論

問題設定

長期間において検証された
Asset Pricing Modelの研究知
見に依拠した問題設定が安心

問題の解き方

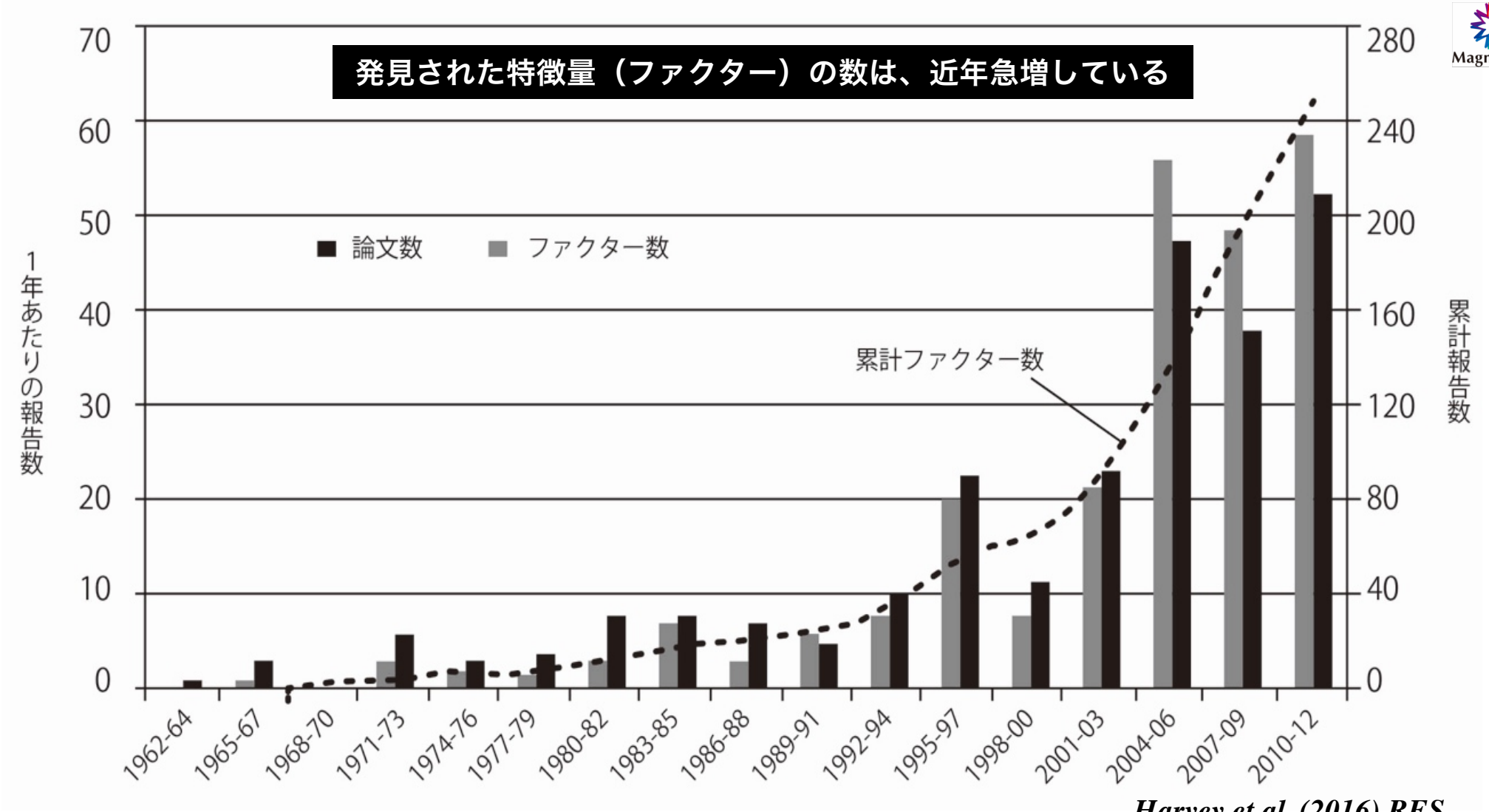
次元の呪いに陥らないように、
次元圧縮を行ってから最適化

ビッグデータとAIで探すべきもの

- ファイナンス領域の研究では、クロスセクションの期待リターンに関する特徴量（ファクター）が数多く発見されている
- こうした特徴量は長期の(100年近い）株価データで確認されている
- 機械学習のアプローチを援用しながら特徴量をどう使うかを考え、より良いポートフォリオ構築の方向性を探索するのが有望
- 高次元のデータを扱うことになるため、自ずと方法論は機械学習となる（LASSO やその他のpenalized regression)
- 未知の特徴量を探すことも重要（後述）

クロスセクションの特徴量 (一例)

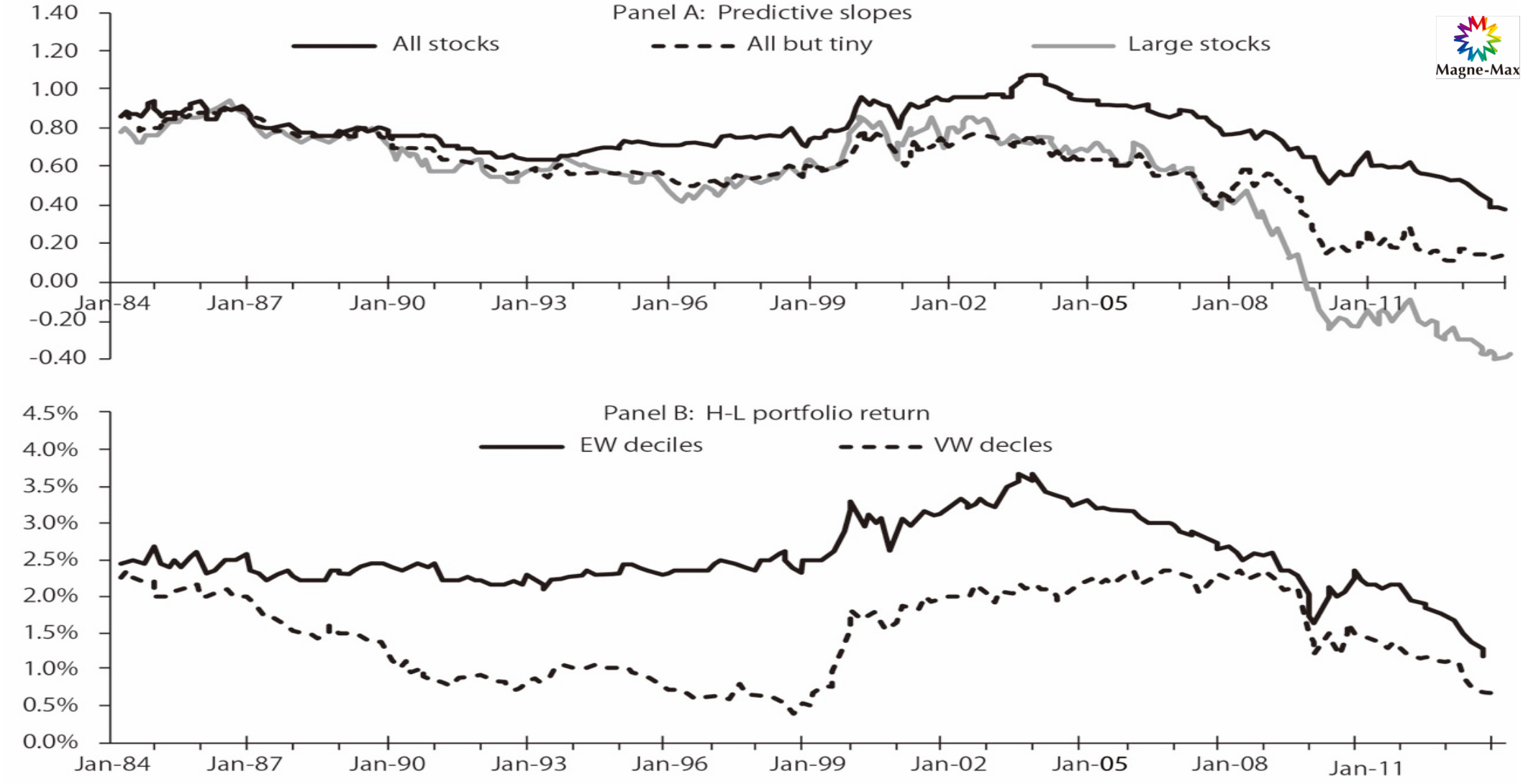
- 会計発生高による期待リターン
- 業績サプライズ
- アナリストによる格付け変更
- アナリストによる業績変更
- 在庫変化率
- 公募増資・自社株買い
- デフォルト確率
- 投資家の注目度
- コーポレート・ガバナンス
- Idiosyncratic volatility



Harvey et al. (2016) RFS

ファイナンス研究と実務応用のジレンマ

- マーケットのメカニズムが明らかにされ、多くの実証研究で特徴量が明らかにされればされるほど、その特徴量を活用しようとする投資家が現れる。
- その結果、その特徴量によって得られる超過リターンは減衰、あるいは、消滅する

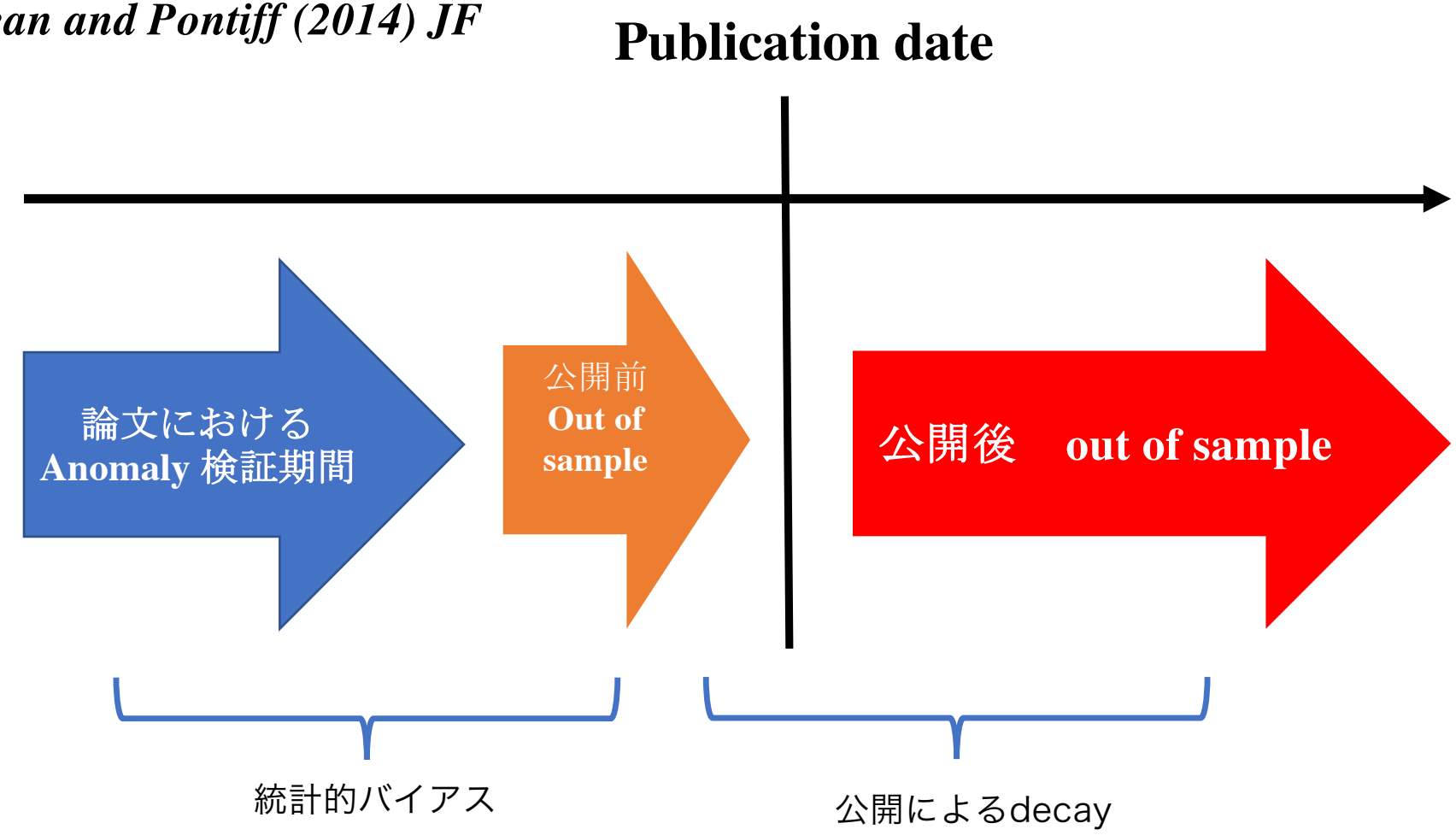


Magne-Max

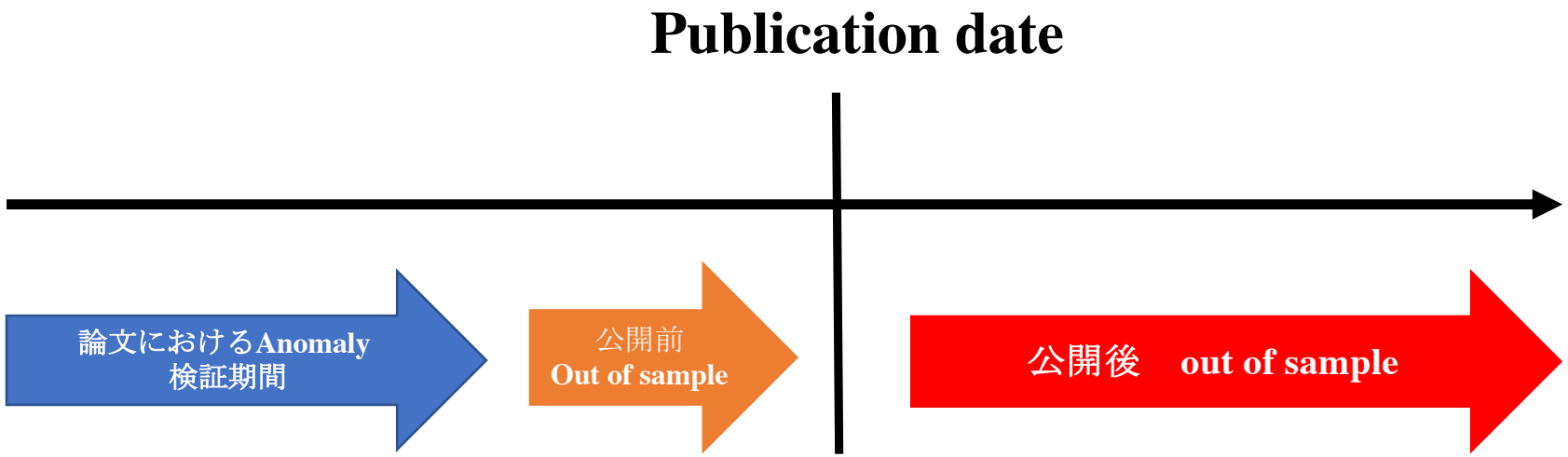
Lewellen (2015) Critical Finance Review

82の特徴量の検証

McLean and Pontiff (2014) JF



McLean and Pontiff (2014) JF



- Out of sample & pre-publicationでは統計的バイアスでabnormal return10%減少
- Out of sample & post-publicationでは平均してabnormal return35%減少
- Out of sample & post-publicationでは、pre-publicationと比較して対象となる分位ポートフォリオの「売買代金」「ボラティリティ」「short interest」「回転率」の全ての指標で上昇
- 既に発表されたanomaly returnと新たにpublish されたanomaly returnの共分散は上昇
- 未発表のanomaly returnとの共分散は下落

まとめ

- 昨今のAIファンドでは、AIによる時系列株価予測が含まれているものが多いようだが、再考の余地はあるのではないか
- ビッグデータの活用ありきではなく、長期間持続している特徴量を用いることが、過学習を防ぐために必要
- 汎化誤差が少ないモデル構築ができても、競争的なマーケットの性質上、普遍性が未来も担保される確証はない
- 新しい特徴量の探索が、モデルの安定化には不可欠