# 競馬予測の苦楽

AlphaImpact
NUKUI Shun

@ PyData.Tokyo One-day Conference 2018

# 自己紹介

- 貫井 駿 (NUKUI Shun, @heartz2001)
- 仕事
  - Fringe81株式会社 (Ad Tech, HR Tech)
  - 競馬予想家
- Python歴: 4年
- 競馬歴: 12年 (馬券は二十歳から)
  - 東工大競馬研究会元部長
- PyData.Tokyo登壇3度目
  - #11 競馬予測 (2016/11)
  - #18 データ分析コンペティションの勝ち上がり方 (2018/04)

第１回電脳賞（春）出場当時

# AlphaImpact

- 競馬AI開発プロジェクト
- メンバー
  - NUKUI Shun: 機械学習、競馬ドメイン
  - OMOTO Tsukasa: 機械学習、システム設計、LightGBMコミッター
  - HARA Tomonori: 競馬ハッカー
- 活動
  - https://alphaimpact.jp/
  - netkeiba.com『ウマい馬券』で予想販売
    - 中央: http://yoso.netkeiba.com/
    - 地方: http://yoso.netkeiba.com/nar/
  - **競馬専門紙『馬サブロー』で予想掲載**

的中型人工知能

回収型人工知能

# 競馬新聞デビュー！

# アジェンダ

- 競馬予測の楽しみ
- 競馬予測の苦しみ
    - 目的変数の選び
    - データ成形
    - 予測モデルの区分決め
    - 予測モデルの学習
    - 予測モデルの評価
    - 売れる馬券の探し方

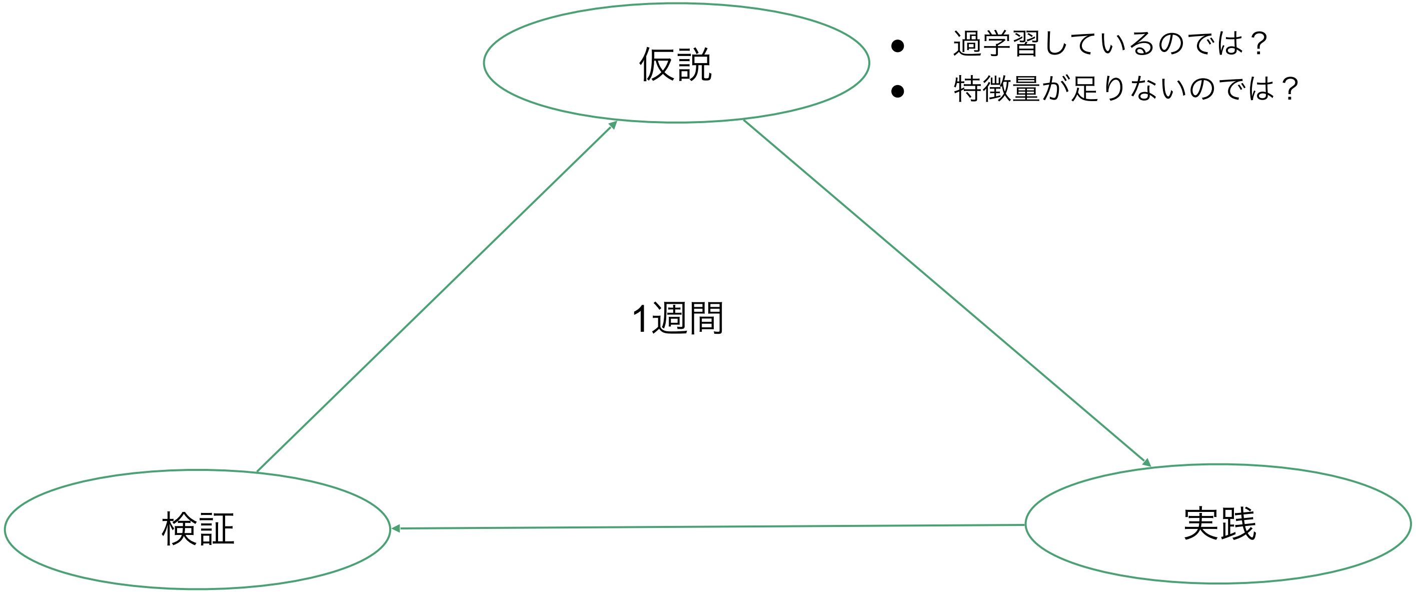
# アジェンダ

- 競馬予測の楽しみ
- 競馬予測の苦しみ
    - 目的変数の選び
    - データ成形
    - 予測モデルの区分決め
    - 予測モデルの学習
    - 予測モデルの評価
    - 売れる馬券の探し方

# 競馬予測が楽しい５つの理由

- 毎週（毎日）解くべき問題となるデータが新しく追加される
- 結果が出る過程をリアルタイムに映像で観ることができる
- 豊富なデータ
- 絶妙な難易度
- 成果がすぐに目に見える形（お金）になる

# 毎週データが追加される




- 過学習しているのでは？
- 特徴量が足りないのでは？

- 運用モデルのパフォーマンス確認
- レースを見る

- 特徴量追加
- 学習方法の変更

# リアルタイムに映像で

# 競馬のデータは豊富

- 構造化データ
  - レース情報、競走成績、騎手、調教師、オッズ、調教など
- テキストデータ
  - 掲示板/Twitter
- 時系列データ
  - 過去戦歴
  - 時系列オッズ
- 映像データ
  - レース映像

# 日本競馬の難易度

- 胴元に売上の20~25%が入るようにオッズが動的に変化する
    - 海外競馬だとブックメーカーが職人芸で固定オッズを決定する
    - パチンコは15%程度
- ランダムに購入した場合、回収率の期待値は75~80%程度
- 回収率100%の壁を超えることは難しいが、不可能ではない

# 成果が目に見える

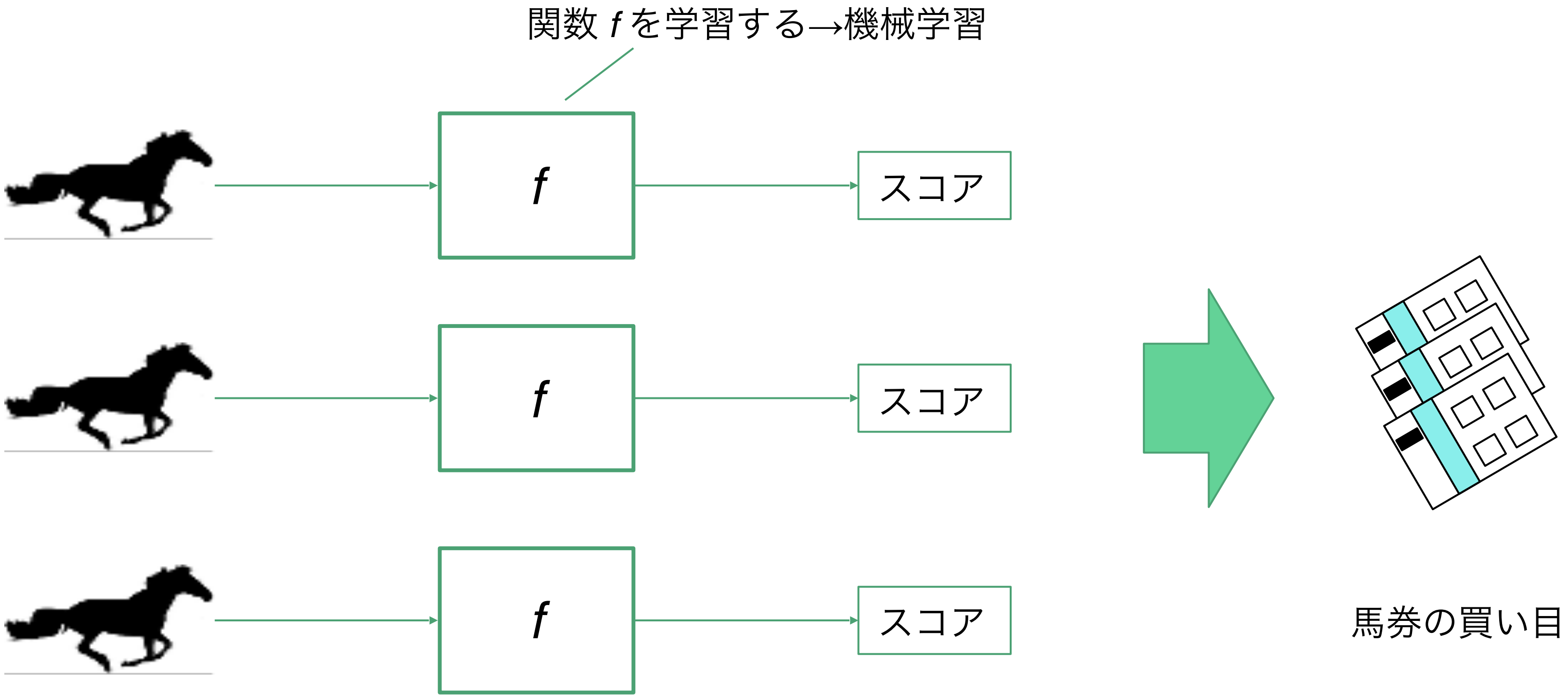
リンク: [http://yoso.netkeiba.com/?pid=yoso_detail&id=506711](http://yoso.netkeiba.com/?pid=yoso_detail&id=506711)

そもそも競馬自体が楽しい

# アジェンダ

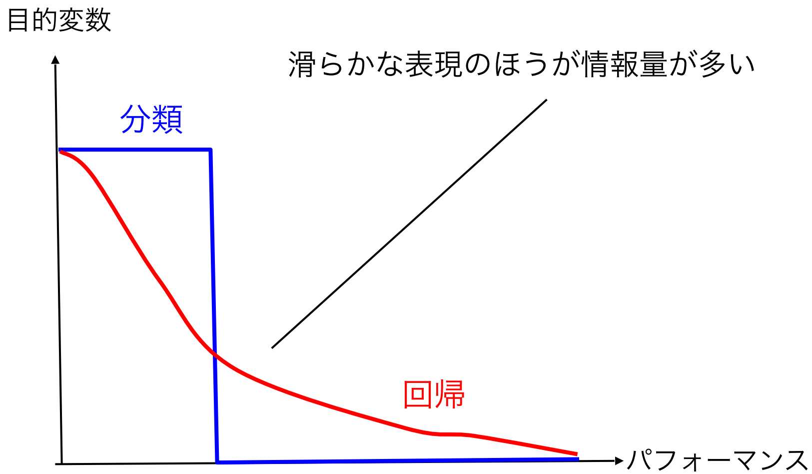
- 競馬予測の楽しみ
- 競馬予測の苦しみ
  - 目的変数の選び
  - データ成形
  - 予測モデルの区分決め
  - 予測モデルの学習
  - 予測モデルの評価
  - 売れる馬券の探し方

# 問題設定



# 何を解くのか？

- 分類問題
  - 0/1に丸めると情報が欠損する
  - ラベル不均衡
- **回帰問題**
  - 自由度が高いので選択のセンスが必要
- ランキング問題
  - 自然なアプローチだが、回帰よりも学習が重くて精度はそこまででもない



# 回帰問題の目的変数

基本となる目的変数

- 着順
- 走破タイム
- 1着からの秒差
- 本賞金
- スピード指数系
- 払戻オッズ

# 目的変数のエンジニアリング

- 特徴量作成やハイパーパラメータ調整よりも重要
- 試行錯誤を繰り返して見つけていくしかない
    - 30種類以上は試した
- ドメイン特化のエンジニアリング
    - レース内標準化
    - 馬券外の馬のスコアは同じにする
    - 着差0.0秒は同着とみなす

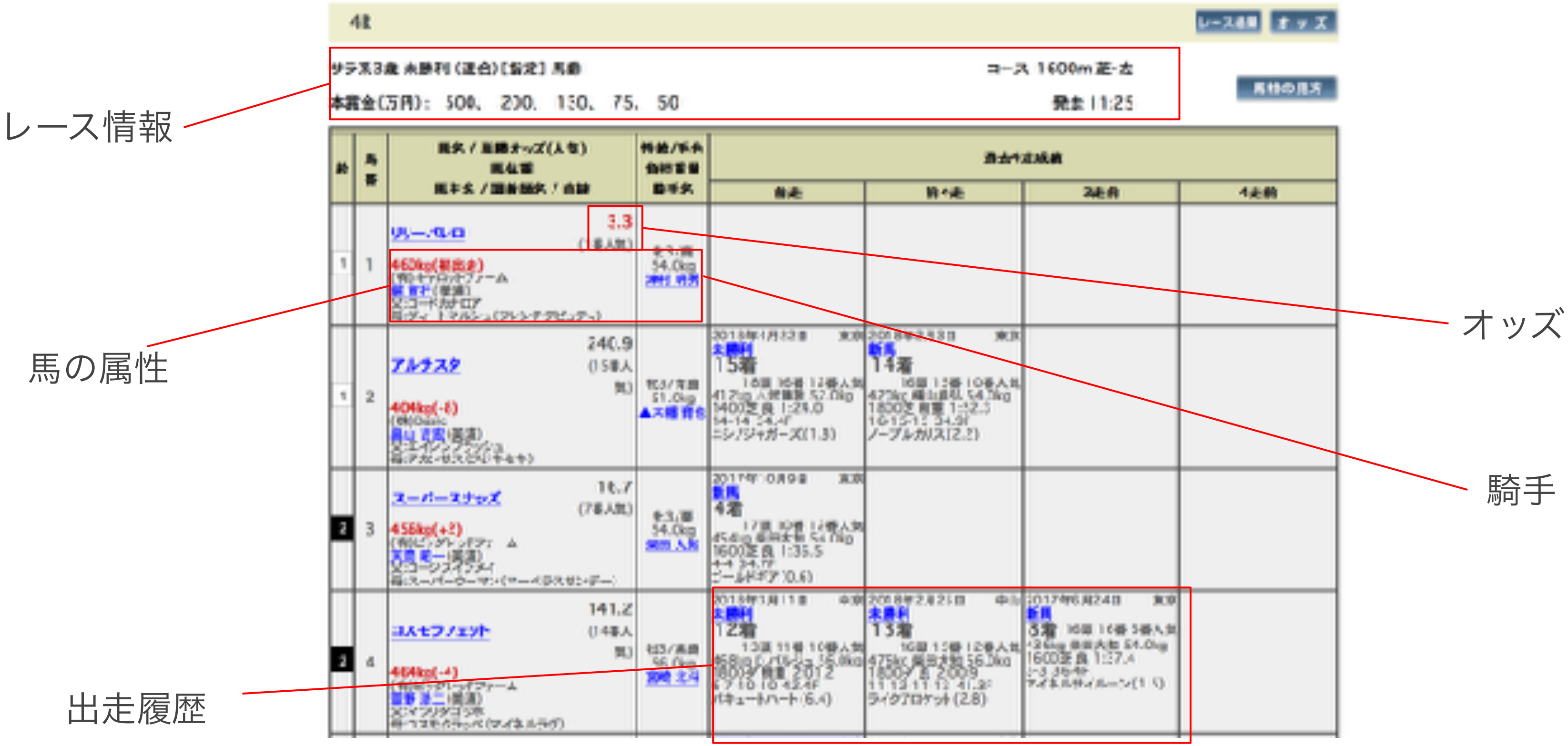
# 解く問題は1つである必要は無い

- 問題を切り分けることにより、解きやすくなる
  - 強さ
  - 美味しさ
  - 人気
- ただし、予測モデルをたくさん作る必要ある

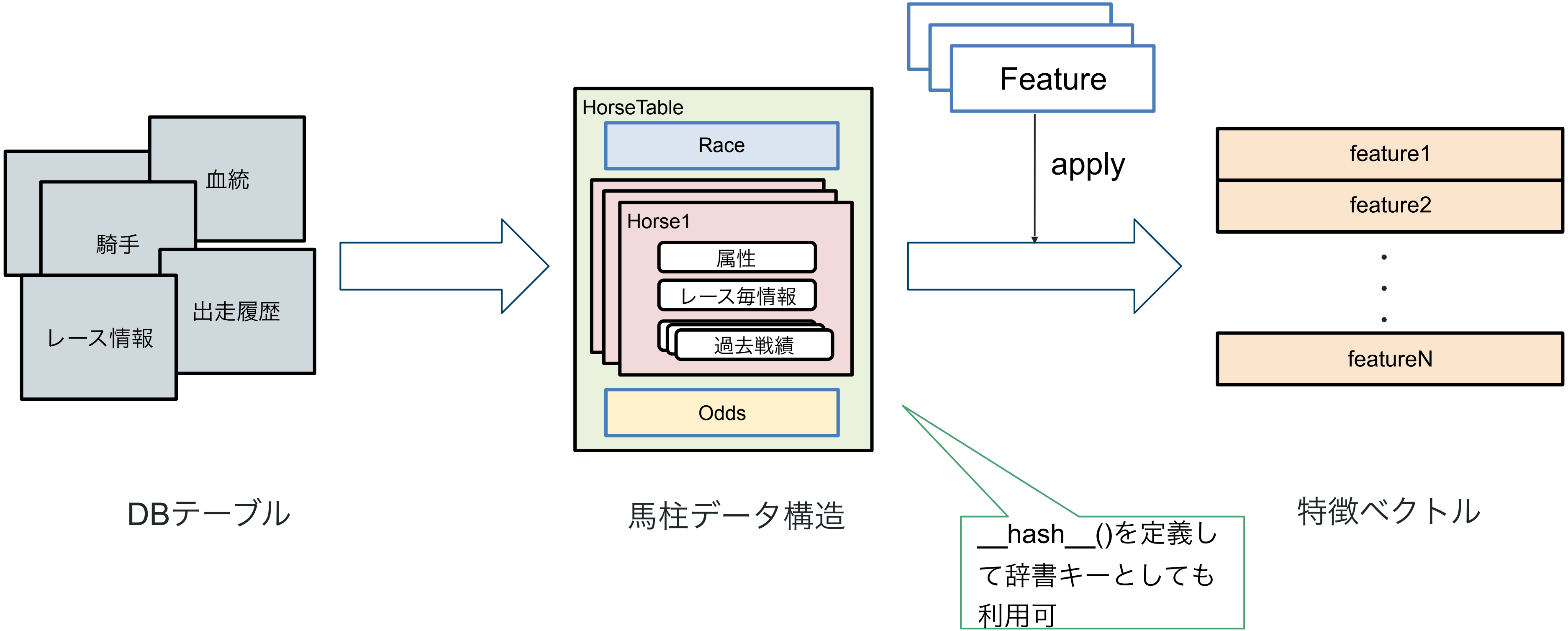
# アジェンダ

- 競馬予測の楽しみ
- 競馬予測の苦しみ
    - 目的変数の選び
    - データ成形
    - 予測モデルの区分決め
    - 予測モデルの学習
    - 予測モデルの評価
    - 売れる馬券の探し方

# 馬柱



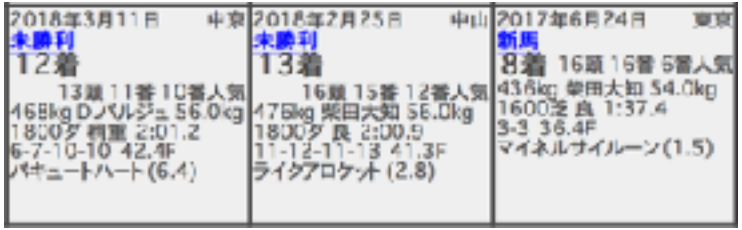
# 馬柱をモデル化してデータ処理を簡潔に



# 出走履歴

- 過去N走の成績を横に並べる
  - カテゴリ変数が大量にあるとNが+1しただけで次元数が爆発的に増える
  - 出走数は競走馬ごとに異なるので、欠損値が多く発生する
  - 次元数の割に情報量が少ない
- 集計値に丸める
  - 次元数を抑えて、情報密度が高い
- RNNのような可変長な入力をとれるモデルを選択するアプローチも考えられる

出走履歴例



| | | |
|---|---|---|
| 2018年3月11日 中京 | 2018年2月25日 中山 | 2017年6月24日 東京 |
| 未勝利 | 未勝利 | 新馬 |
| 12着 | 13着 | 8着 16頭 16番 5番人気 |
| 13頭 11番 10番人気 | 16頭 15番 12番人気 | 436kg 柴田大知 54.0kg |
| 468kg Dバルジュ 56.0kg | 476kg 柴田大知 56.0kg | 1600芝 良 1:37.4 |
| 1800ダ 稍重 2:01.2 | 1800ダ 良 2:00.9 | 3-3 36.4F |
| 6-7-10-10 42.4F | 11-12-11-13 41.3F | マイネルサイルーン(1.5) |
| パキュートハート(6.4) | ライタアロケット(2.8) | |

# 出走履歴の集計特徴量

- 「レース条件」の切り分け方は競馬のドメイン知識が必要
- 組み合わせ的に次元数が増えるので、計算コストは増加する
  - 2000個以上を作成

カテゴリ
- 馬
- 騎手
- 馬×騎手
- 調教師
- 馬主
- 父馬
- …

×

レース条件
- 芝・ダート・障害
- 競馬場
- 距離
- 競馬場×距離
- 天候
- 馬場状態
- ペース
- …

×

統計値
- 出走回数
- 勝率
- 複勝率
- 単勝回収率
- 複勝回収率
- …

# オッズ

- 的中したときの賭け金に対する払戻し倍率
- オッズは馬券発売締切まで刻一刻と変動する
- 過去データで学習したときに確定オッズを特徴量として使うと、シミュレーションの性能と実運用したときの性能に乖離が起きる
- 前日予測では使えない
    - 発売されていても確定オッズと乖離が大きすぎる
- 時系列オッズ解析だけで馬券を選択するアプローチも可

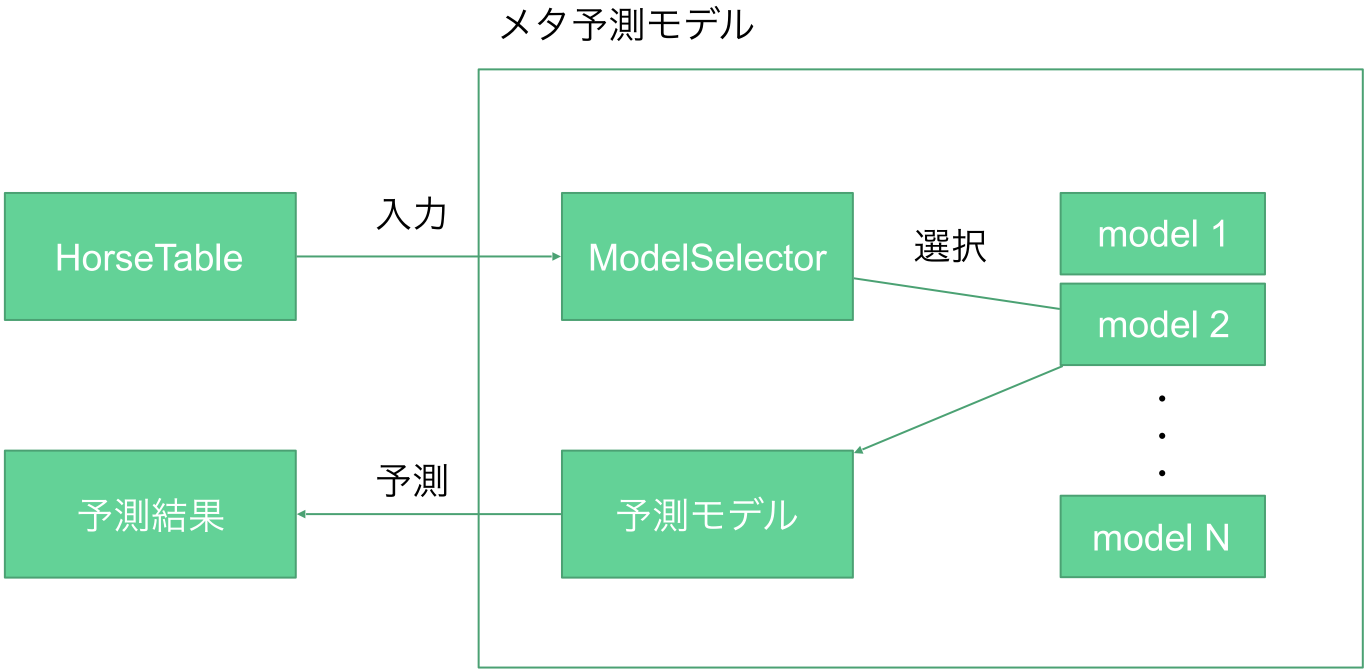
# アジェンダ

- 競馬予測の楽しみ
- 競馬予測の苦しみ
  - 目的変数の選び
  - データ成形
  - 予測モデルの区分決め
  - 予測モデルの学習
  - 予測モデルの評価
  - 売れる馬券の探し方

# レース区分

- 競馬場
- 芝・ダート・障害
- 距離
- 内回り・外回り
- クラス条件（新馬、未勝利、1600万下など）

# 適切なレース区分ごとに予測モデルを選択

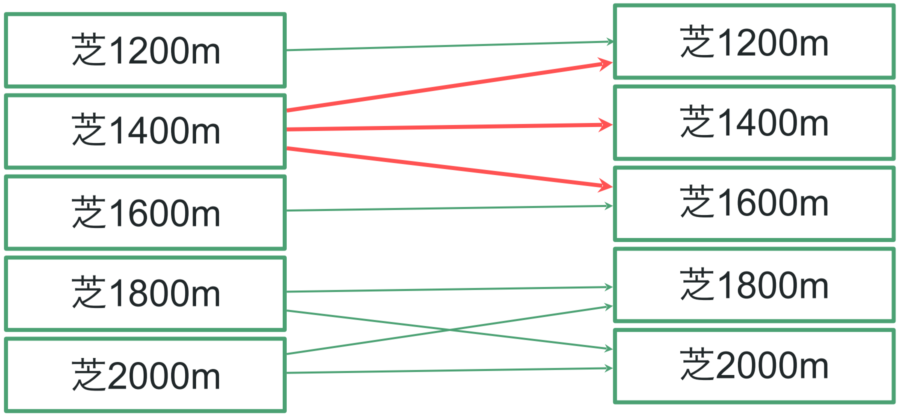


# 適切なレース区分ごとに予測モデルを作成

- どの粒度が適切かは、**特徴量**と**目的変数**に大きく依存する
  - 走破タイムはコースに依存する
  - オッズはコースに依存しない
- 学習データと予測対象データの区分は必ずしも一致する必要は無い

例)

予測データ区分　　　　学習データ区分



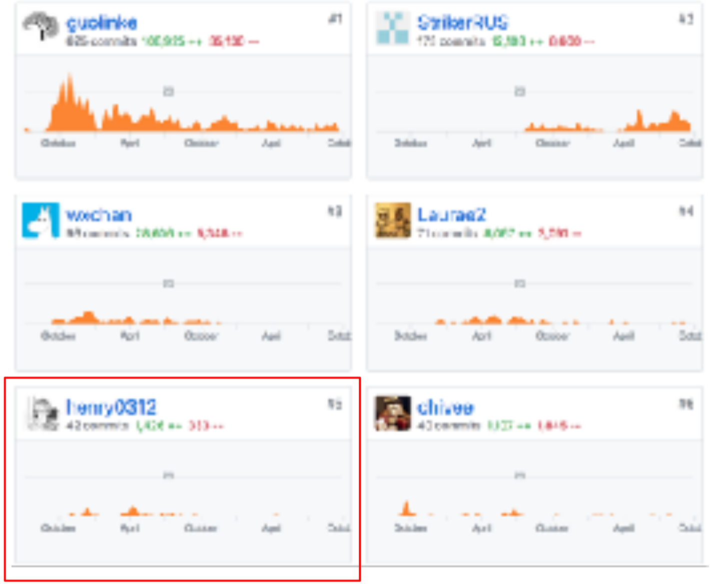

芝1400m予測モデルは芝1200-1600mで学習する

# アジェンダ

- 競馬予測の楽しみ
- 競馬予測の苦しみ
  - 目的変数の選び
  - データ成形
  - 予測モデルの区分決め
  - 予測モデルの学習
  - 予測モデルの評価
  - 売れる馬券の探し方

# LightGBM

- 勾配ブースティングの高速・軽量・高精度な実装
- カテゴリ変数をカテゴリ変数として扱うことができる
- 欠損値を欠損値として扱うことができる
- AlphaImpactのOMOTOはLightGBMのcommiter
- https://github.com/Microsoft/LightGBM



# ハイパーパラメータのチューニング

- hyperoptを利用
  - ベイズ最適化
  - https://github.com/hyperopt/hyperopt
- 5-fold cross validation
- 探索パラメータ数が多いので、計算コストが高い
- 200回ほど探索

# 予測モデルの実験

- 目的変数
- 特徴量
- モデル区分
- ハイパーパラメータ

試すべき条件が多く計算時間が非常にかかる

# GCEのプリエンプティブインスタンスの活用

- Google Compute Engineの計算余剰資源を利用しているため、利用料金が約70%オフ
- 1 プロジェクト 72CPUまでなので、それ以上必要な場合はプロジェクトを分ける必要がある
  - プロジェクト間でイメージのコピーはできない
- 予告なしに落ちる（最大24時間）なので、探索状態を逐次保存しておき、再開できるようにしておく
  - hyperopt.Trialsをpickle保存しておけばOK
- 実験の進捗は逐次Slackに通知

# アジェンダ

- 競馬予測の楽しみ
- 競馬予測の苦しみ
  - 目的変数の選び
  - データ成形
  - 予測モデルの区分決め
  - 予測モデルの学習
  - 予測モデルの評価
  - 売れる馬券の探し方

# nDCG

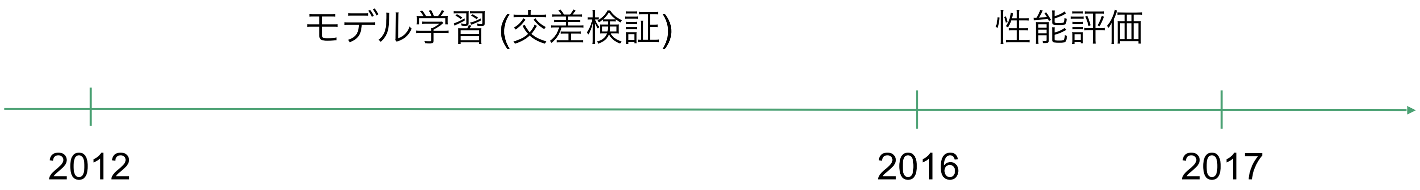
- ランキング問題でよく使われるnDCGを利用
- 高い関連度がより上位に予測できていれば大きな値になる（最大値1）
- 関連度は自分で定義する必要あり
  - 非負値かつ大きい＝良いとなる数値
  - 賞金がおすすめ

$$DCG = rel_1 + \sum_{i=2}^{n} \frac{rel_i}{\log_2 i}$$

$$nDCG = \frac{DCG}{idealDCG}$$

# 評価データの期間

- 競馬のデータは少なくて貴重
  - 年間３千レース程度
- 季節性なども考慮して最低1年間は評価に必要
- コースの改修があるので、あまりに古いデータや役に立たない



# コースの改修工事

- 競馬場は不定期に改修が行われる
- 大改修の場合は傾向が大きく変わってしまうので、注意が必要
- 最近の主な大改修
  - 2006年 阪神競馬場大改修
  - 2010年4月~2012年1月 中京競馬場大改修

# アジェンダ

- 競馬予測の楽しみ
- 競馬予測の苦しみ
  - 目的変数の選び
  - データ成形
  - 予測モデルの区分決め
  - 予測モデルの学習
  - 予測モデルの評価
  - 売れる馬券の探し方

# 「勝てる馬券」≠「売れる馬券」

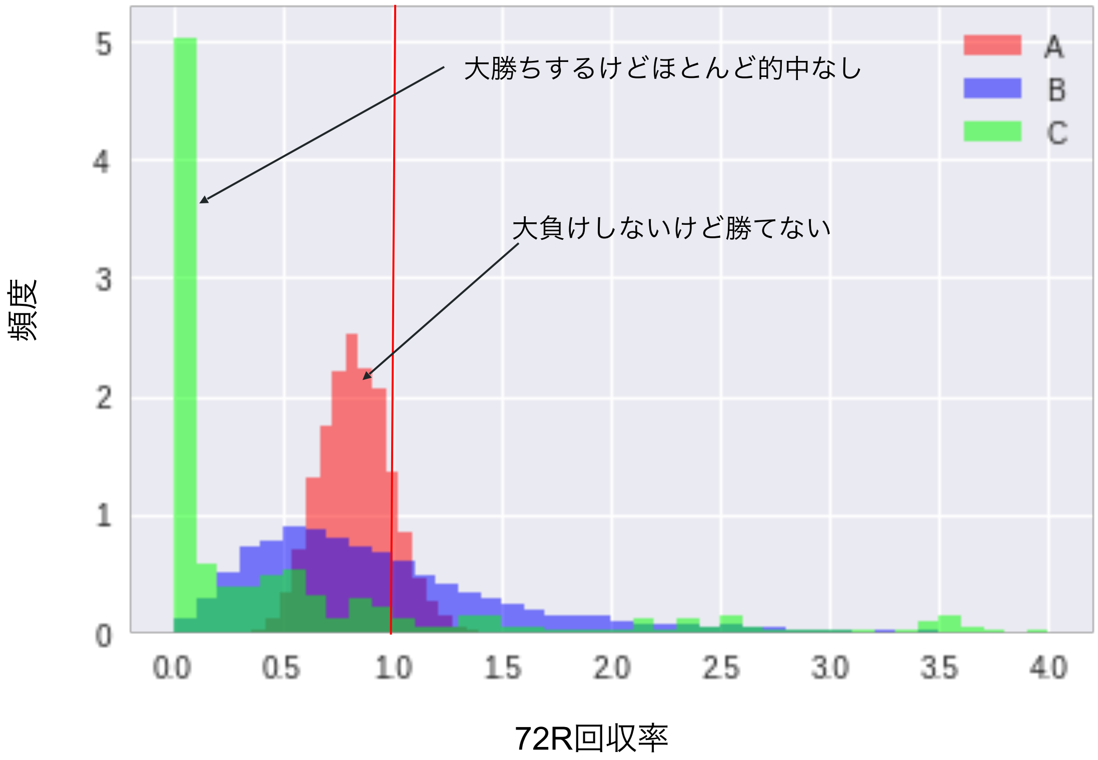
1. 100週間のうち99週間は回収率0%だが、**通算回収率200%**になる馬券
2. 100週間のうち90週間は回収率100%超えだが、**通算回収率105%**になる馬券

⇒ (1) の馬券を買い続けるには強靭な忍耐力が必要＝売れない馬券

# 1週間回収率100%超え確率

- 馬券購入者の気持ちとして、週間単位で安定してプラスになると嬉しい
- 3場開催の場合、1週間に72Rが行われる
- 評価データから72Rを繰り返し復元抽出し、100%超えの確率を計算する

# 1週間回収率100%超え確率



# 馬券の利用性

- ３連単などの購入点数が多くなりやすい馬券種では、ボックスやフォーメーションなどの賭け方で一括購入ができる
- 等価な馬券をバラで購入可能だが、大変面倒くさい

3連単フォーメーション

A, B - A, B, C - A, B, C, D, E 各100円 → 展開

| | |
|---|---|
| A - B - C 100円 | B - A - C 100円 |
| A - B - D 100円 | B - A - D 100円 |
| A - B - E 100円 | B - A - E 100円 |
| A - C - B 100円 | B - C - A 100円 |
| A - B - D 100円 | B - A - D 100円 |
| A - B - E 100円 | B - A - E 100円 |

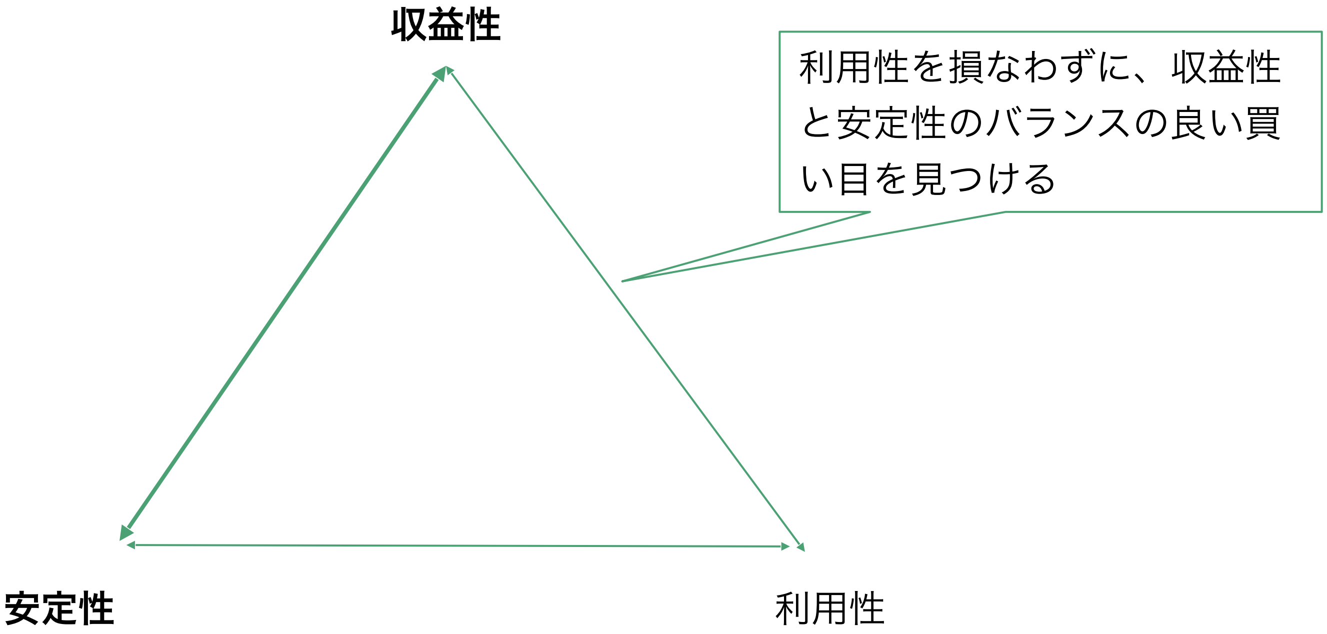
# 馬券の利用性

- フォーメーションで買えないような多点数馬券はユーザにとって買いずらい

| | | | | |
|---|---|---|---|---|
| A - B - C | 800円 | | B - A - C | 200円 |
| A - B - D | 200円 | | B - A - D | 300円 |
| A - B - E | 500円 | ・・・ | B - A - E | 800円 |
| A - C - B | 300円 | | B - C - A | 300円 |
| A - B - D | 100円 | | B - A - D | 1000円 |
| A - B - E | 1000円 | | B - A - E | 900円 |

計120点

# 馬券の３つの評価軸



# まとめ

- 競馬は楽しい、競馬予測はもっと楽しい
- 競馬データはクセがある
- 予測モデル実験は目的変数・特徴量・レース区分・ハイパーパラメータを探索
- 競馬は意外とデータが少なく評価が難しい
- 馬券の良し悪しは「収益性」「安定性」「利用性」で判断する