機械学習で競馬を勝つために必要なこと

AlphaImpact
NUKUI Shun
PyData.Tokyo #11

自己紹介

名前: 貫井 駿 (@heartz2001)

競馬歴: 10年

好きな馬: ハーツクライ

所属: 東京工業大学 情報理工学研究科

　　計算工学専攻

専門: ソーシャルネットワーク・機械学習

サークル: 東工大競馬研究会 部長

競馬活動: AlphaKeiba開発

AlphaImpactプロジェクト

- 第1回ウマナリティクスで出会った大元氏(@henry0312)と共同で最強の競馬AIを開発すべく発足
- ディープラーニングを使っています(言いたかった)
- 6月頃から開発をスタートし、10月より予測を毎週公開中

twitter: @alphaimpact_ai
HP: https://alphaimpact.jp/

Contents

1. データの収集
2. 特徴量の作成
3. モデルの設計
4. 学習データの分割
5. 性能評価

Contents

netkeiba.com[1]

- 無料で使える (有料コンテンツもあり)
- スクレイピングしやすいHTML
- JRAが公式で公開している情報はだいたい網羅されている
- 有料会員(月額500円)に入れば調教データも取得できる
- 掲示板が馬ごとに掲示板があるのでNLPのアプローチでも？

- 『競馬の予測をガチでやってみた』[2]のstockedge氏がnetkeiba-scraperを公開している

GitHub: https://github.com/stockedge/netkeiba-scraper

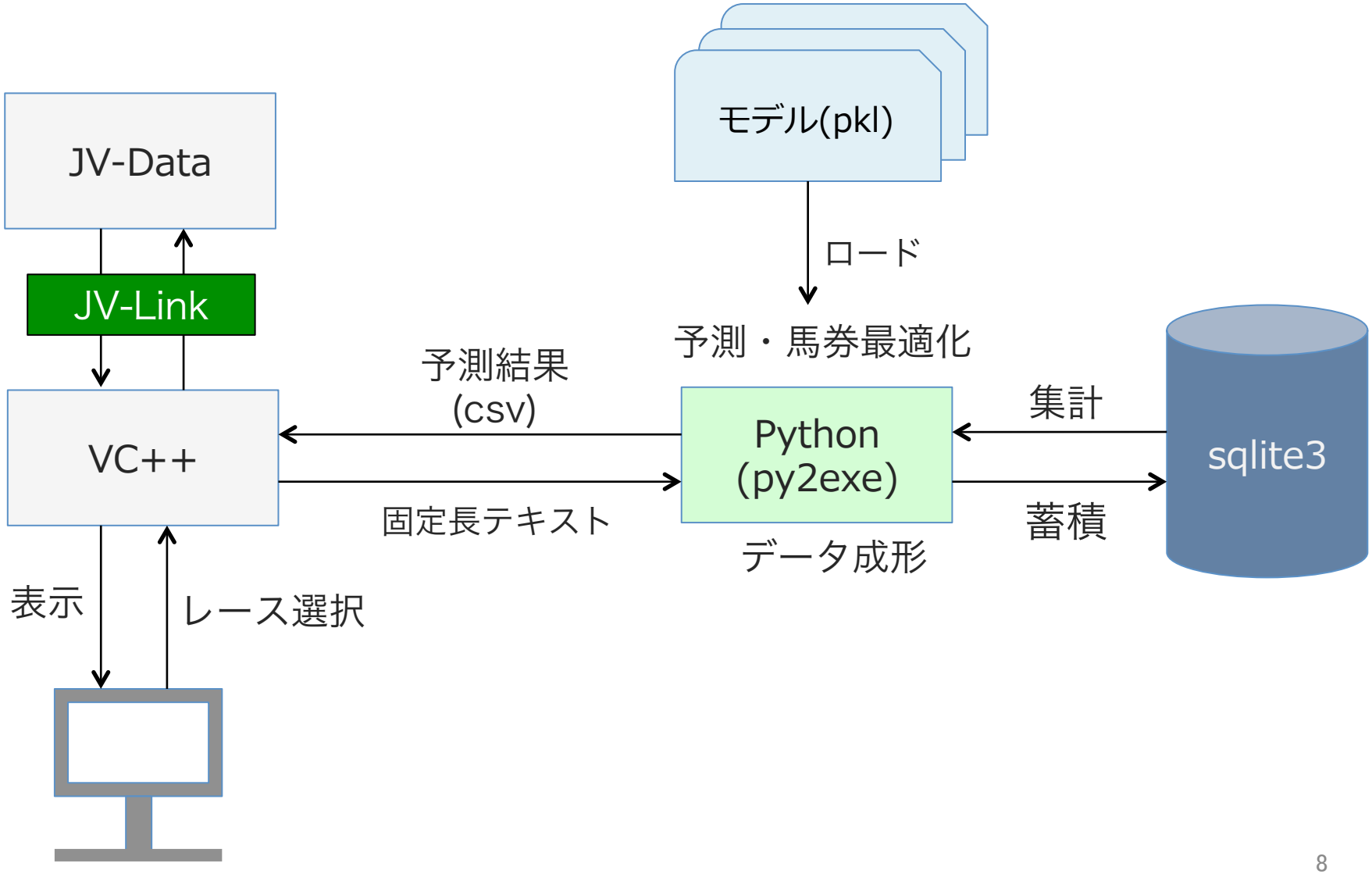
[1] http://www.netkeiba.com/
[2] http://stockedge.hatenablog.com/entry/2016/01/03/103428

JRA-VAN (DataLab.)[3]

- 公式なので信頼できる
- 唯一リアルタイム投票数が取れる(単勝、複勝、馬連のみ)
- 月額 2,052円
- 固定長テキスト形式
- APIがVisual Studio C++/C#にのみ対応(=Windows縛り)
- Python使いとの相性は悪い

[3] http://jra-van.jp/dlb

DataLab.の利用例 (AlphaKeibaの場合)



JRDB[4]

- 古いlzh形式で配布されているが、URLを叩いて解凍すれば固定長テキストが取得できるので、どのOSでも使える
- 99年からデータが提供されている(一部データ除く)
- パドックの生の情報(馬体状態)などの独自コンテンツが豊富
- IDMや基準オッズを始めとする様々な指数を提供しているのでテクニカル分析もできる
- 馬券裁判でおなじみの卍氏も使っている

[4] http://www.jrdb.com/

データソースまとめ

データソース	月額料金	取得方法	メリット	デメリット
JRA HP http://www.jra.go.jp/	無料	スクレイピング	リアルタイムにオッズが取得できる	クローリングしずらい
JRA-VAN (DataLab.) http://jra-van.jp/dlb	2,052円	固定長テキスト	リアルタイムの投票数や坂路調教が取れる.	Visual Studio(Windows)でないと動かせないAPI
netkeiba.com http://www.netkeiba.com/	無料(500円)	スクレイピング	基本的な情報をお手軽に取得できる	情報量はそんなに多くない
JRDB http://www.jrdb.com/	1,980円	固定長テキスト	公式に無い馬場差やパドックなど+αの情報が豊富. lzh形式で配布されているのでLinux/OSXでも使える	お金がかかる

Contents

馬柱

2016年11月20日(日) 3回福島6日

開催選択へ戻る　レース選択へ戻る　印刷用ページ

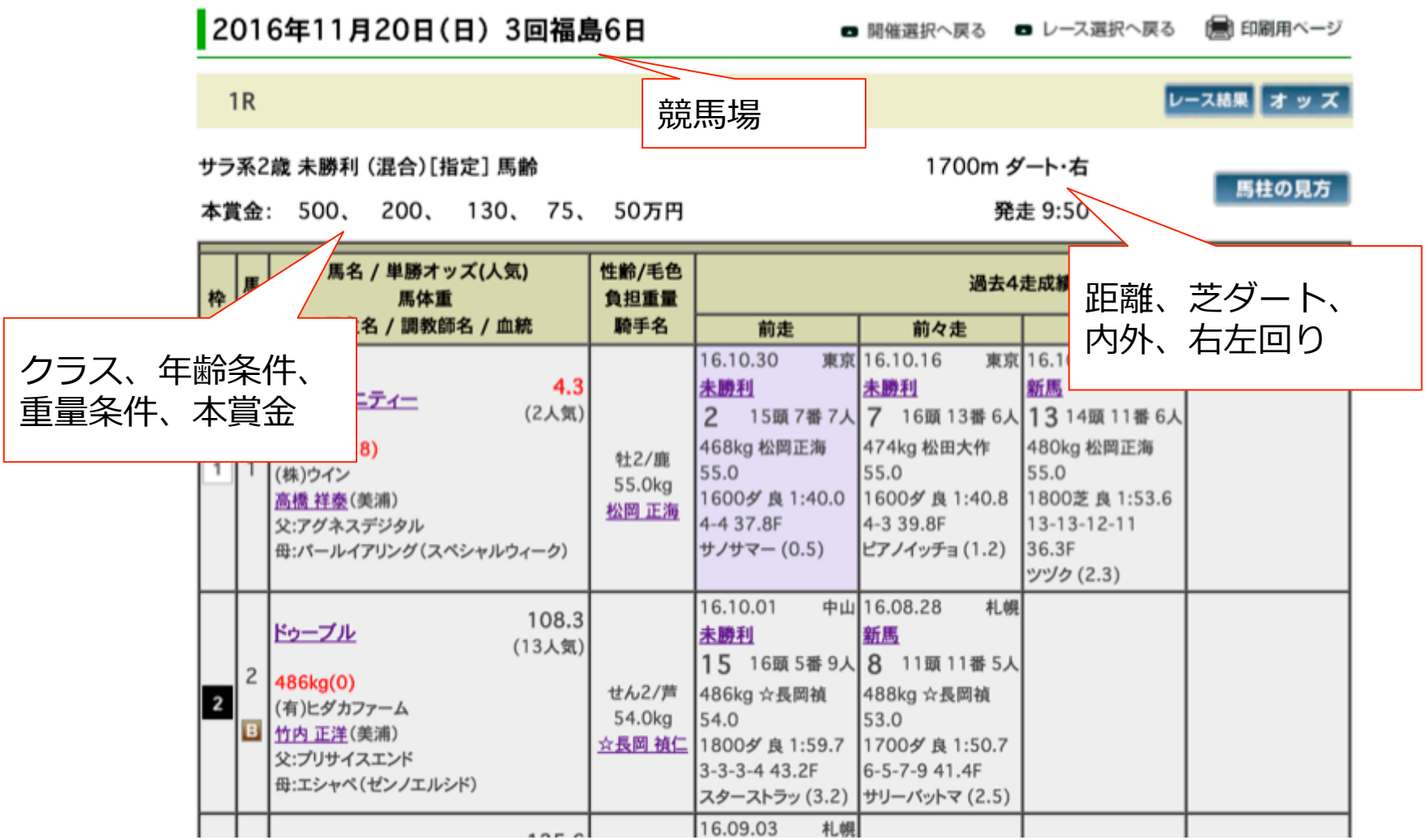
1R　レース結果　オッズ

サラ系2歳 未勝利 (混合)[指定] 馬齢　1700m ダート・右

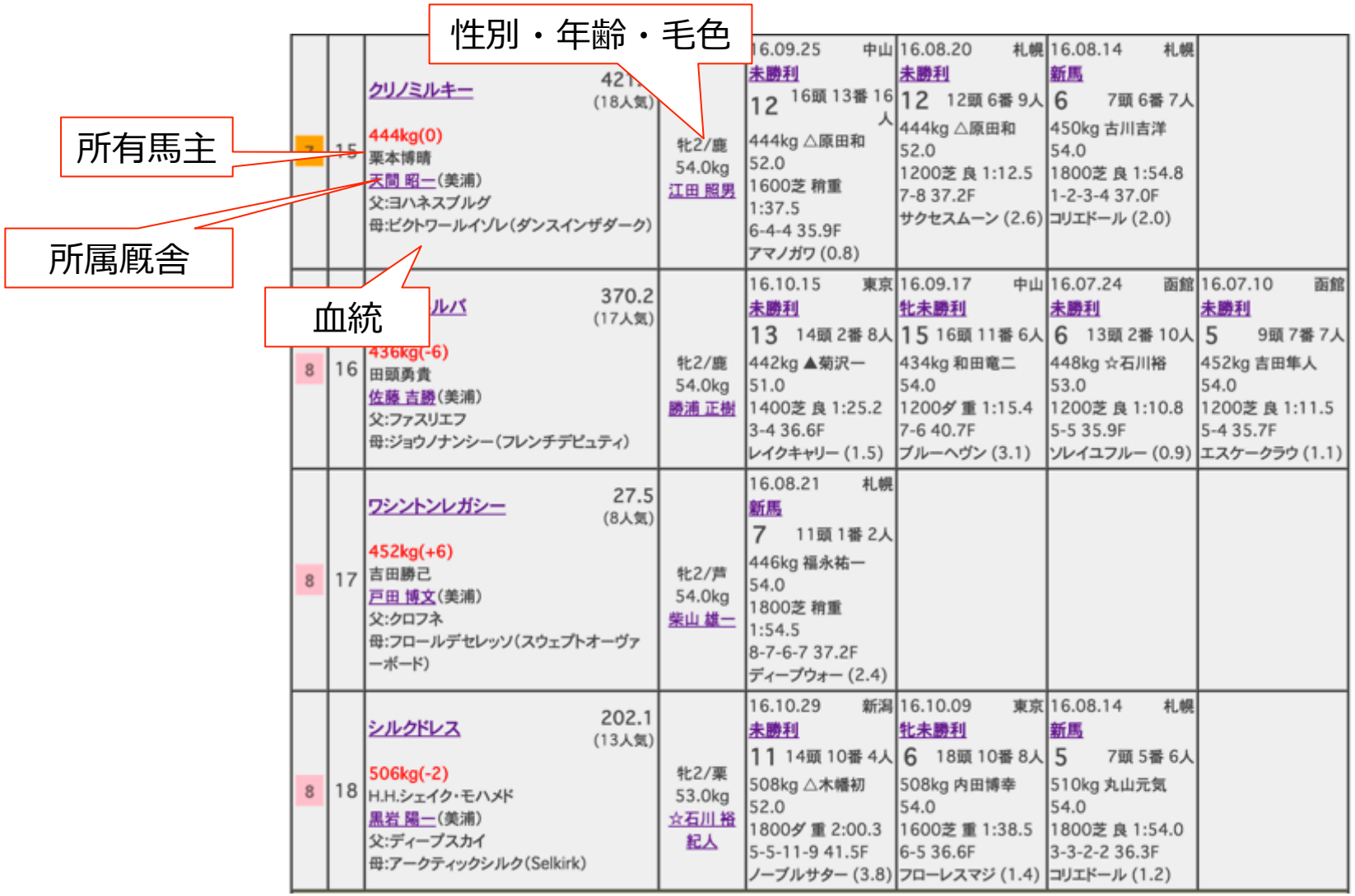
本賞金: 500、 200、 130、 75、 50万円　発走 9:50　馬柱の見方

枠	馬番	馬名 / 単勝オッズ(人気) 馬体重 馬主名 / 調教師名 / 血統	性齢/毛色 負担重量 騎手名	過去4走成績 前走	前々走	3走前	4走前
1	1	ウイントリニティー　4.3 (2人気) 486kg(+18) (株)ウイン 高橋 祥泰(美浦) 父:アグネスデジタル 母:パールイアリング(スペシャルウィーク)	牡2/鹿 55.0kg 松岡 正海	16.10.30 東京 未勝利 2 15頭 7番 7人 468kg 松岡正海 55.0 1600ダ 良 1:40.0 4-4 37.8F サノサマー (0.5)	16.10.16 東京 未勝利 7 16頭 13番 6人 474kg 松田大作 55.0 1600ダ 良 1:40.8 4-3 39.8F ピアノイッチョ (1.2)	16.10.02 中山 新馬 13 14頭 11番 6人 480kg 松岡正海 55.0 1800芝 良 1:53.6 13-13-12-11 36.3F ツヅク (2.3)	
2	2 B	ドゥーブル　108.3 (13人気) 486kg(0) (有)ヒダカファーム 竹内 正洋(美浦) 父:ブリサイスエンド 母:エシャペ(ゼンノエルシド)	せん2/芦 54.0kg ☆長岡 禎仁	16.10.01 中山 未勝利 15 16頭 5番 9人 486kg ☆長岡禎 54.0 1800ダ 良 1:59.7 3-3-3-4 43.2F スタートラッ (3.2)	16.08.28 札幌 新馬 8 11頭 11番 5人 488kg ☆長岡禎 53.0 1700ダ 良 1:50.7 6-5-7-9 41.4F サリーバットマ (2.5)		
2	3	キモンズラブ　135.6 (14人気) 454kg(0) 小林祥晃 武藤 善則(美浦) 父:キモンノカシワ 母:スパイオブラヴ(フレンチデピュティ)	牝2/芦 54.0kg 西田 雄一郎	16.09.03 札幌 新馬 7 10頭 3番 9人 454kg 横山和生 54.0 1500芝 良 1:34.3 8-9-7 36.2F ヤマカツグレー (1.8)			
3	4	スプリングフット　2.6 (1人気) 474kg(+2) 島川隆哉 小檜山 悟(美浦) 父:トーセンブライト 母:トーセンパピヨン(フジキセキ)	牡2/鹿 55.0kg 横山 和生	16.11.12 福島 未勝利 2 15頭 13番 3人 472kg 宮崎北斗 55.0 1700ダ 不良 1:47.0 6-5-2-2 39.2F カンムル (0.0)	16.10.08 東京 未勝利 3 16頭 15番 8人 458kg 宮崎北斗 55.0 1400ダ 稍重 1:26.9 5-4 37.9F レッドオーガー (0.9)	16.09.18 中山 未勝利 5 16頭 1番 16人 458kg 宮崎北斗 54.0 1800ダ 稍重 1:57.6 4-4-9-8 40.3F マイネルレンカ (1.6)	16.08.06 新潟 未勝利 11 11頭 7番 11人 462kg 宮崎北斗 54.0 1800芝 良 1:51.5 10-10 36.5F レジェンドセラ (3.7)

馬柱(レース情報)

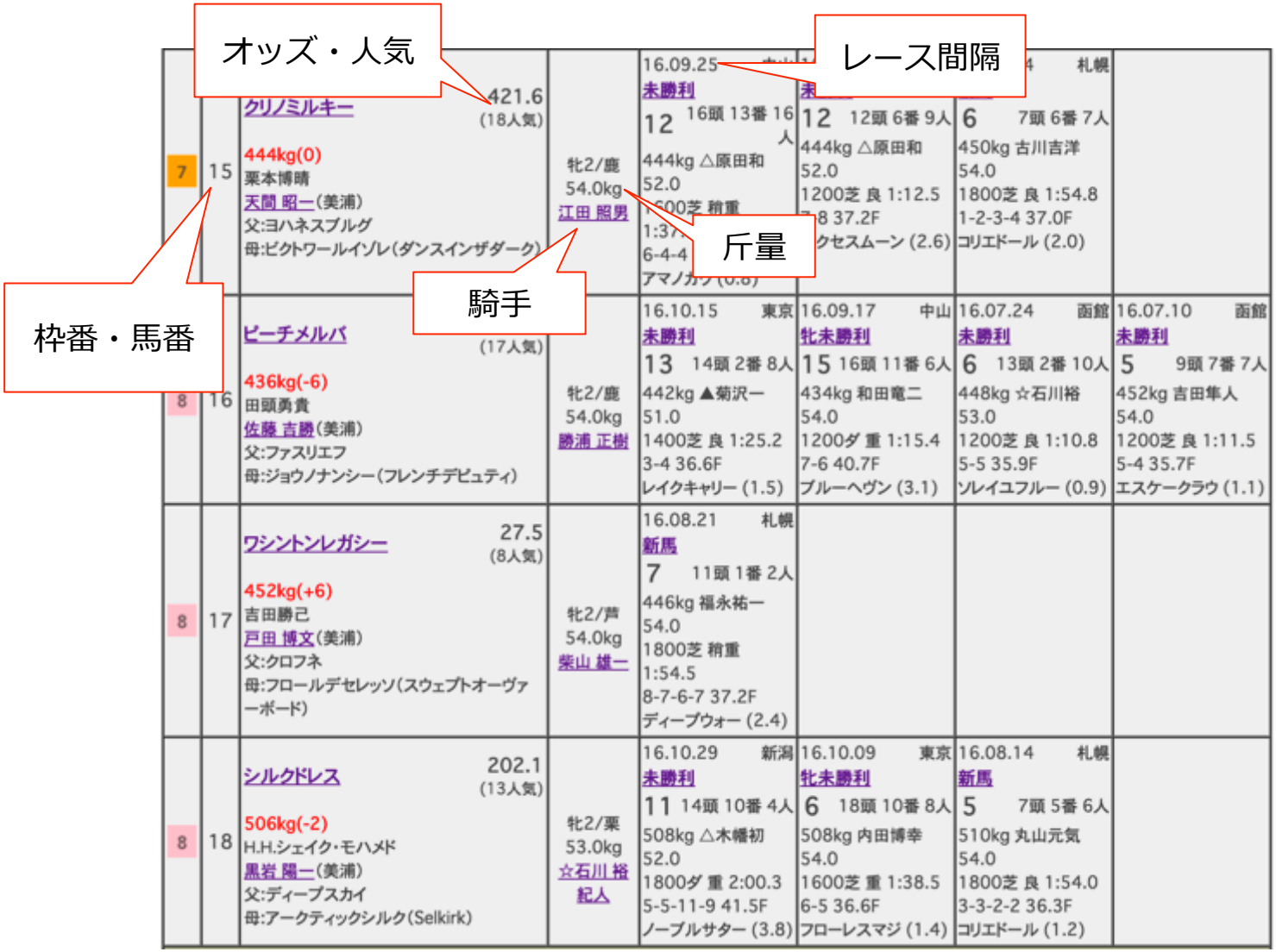

2016年11月20日(日) 3回福島6日
開催選択へ戻る
レース選択へ戻る
印刷用ページ
1R
競馬場
レース結果
オッズ
サラ系2歳 未勝利 (混合)[指定] 馬齢
1700m ダート・右
本賞金: 500、 200、 130、 75、 50万円
発走 9:50
馬柱の見方
クラス、年齢条件、
重量条件、本賞金
距離、芝ダート、
内外、右左回り
枠
馬名 / 単勝オッズ(人気)
馬体重
名 / 調教師名 / 血統
性齢/毛色
負担重量
騎手名
過去4走成績
前走
前々走
4.3
(2人気)
(株)ウイン
高橋 祥泰(美浦)
父:アグネスデジタル
母:パールイアリング(スペシャルウィーク)
牡2/鹿
55.0kg
松岡 正海
16.10.30 東京
未勝利
2 15頭 7番 7人
468kg 松岡正海
55.0
1600ダ 良 1:40.0
4-4 37.8F
サノサマー (0.5)
16.10.16 東京
未勝利
7 16頭 13番 6人
474kg 松田大作
55.0
1600ダ 良 1:40.8
4-3 39.8F
ピアノイッチョ (1.2)
新馬
13 14頭 11番 6人
480kg 松岡正海
55.0
1800芝 良 1:53.6
13-13-12-11
36.3F
ツヅク (2.3)
2
ドゥーブル
108.3
(13人気)
486kg(0)
(有)ヒダカファーム
竹内 正洋(美浦)
父:プリサイスエンド
母:エシャペ(ゼンノエルシド)
せん2/芦
54.0kg
☆長岡 禎仁
16.10.01 中山
未勝利
15 16頭 5番 9人
486kg ☆長岡禎
54.0
1800ダ 良 1:59.7
3-3-3-4 43.2F
スターストラッ (3.2)
16.08.28 札幌
新馬
8 11頭 11番 5人
488kg ☆長岡禎
53.0
1700ダ 良 1:50.7
6-5-7-9 41.4F
サリーバットマ (2.5)
16.09.03 札幌

馬柱 (馬の属性)



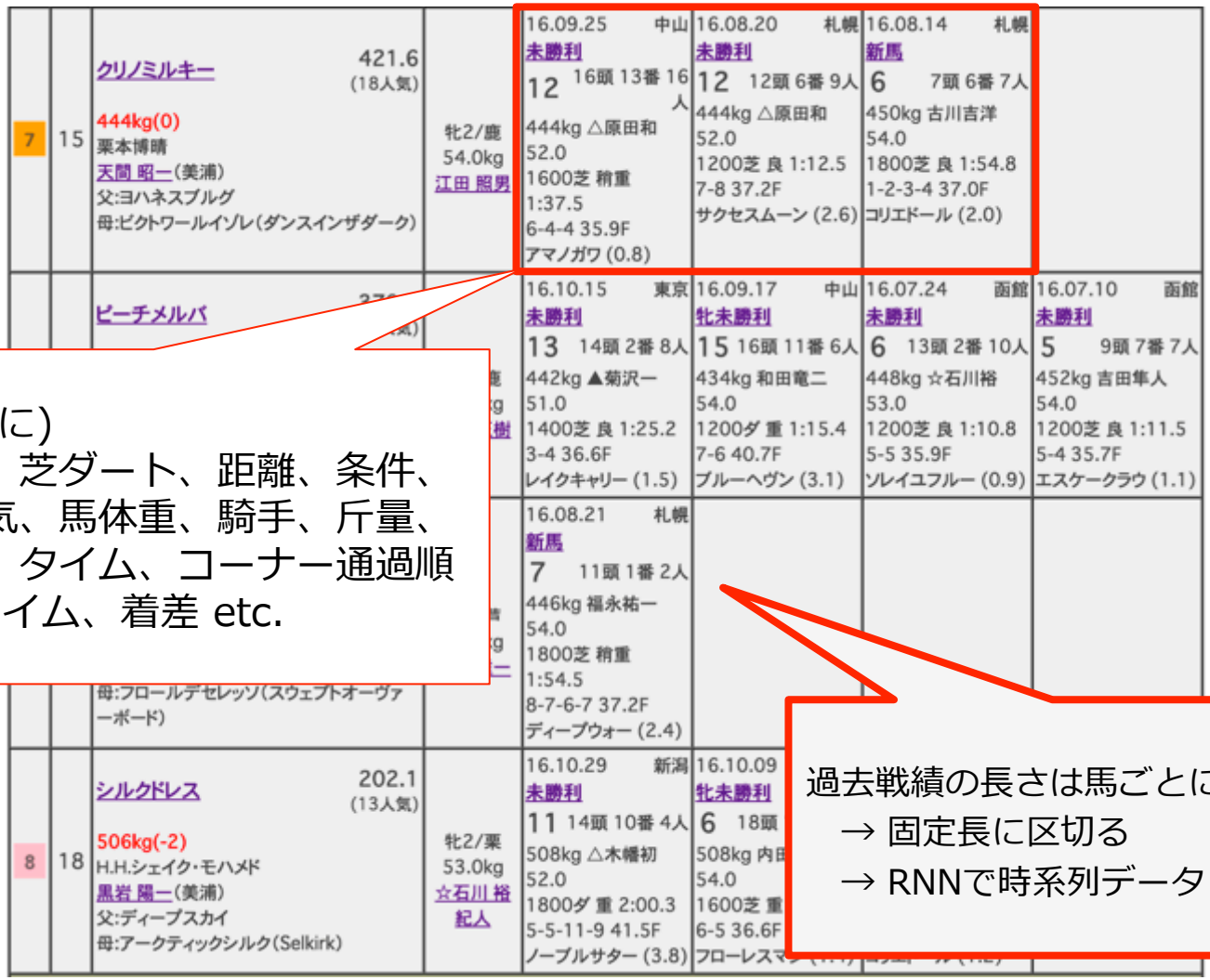
7	15	クリノミルキー 444kg(0) 栗本博晴 天間 昭一(美浦) 父:ヨハネスブルグ 母:ビクトワールイゾレ(ダンスインザダーク)	421. (18人気)	牝2/鹿 54.0kg 江田 照男	16.09.25 中山 未勝利 12 16頭 13番 16人 444kg △原田和 52.0 1600芝 稍重 1:37.5 6-4-4 35.9F アマノガワ (0.8)	16.08.20 札幌 未勝利 12 12頭 6番 9人 444kg △原田和 52.0 1200芝 良 1:12.5 7-8 37.2F サクセスムーン (2.6)	16.08.14 札幌 新馬 6 7頭 6番 7人 450kg 古川吉洋 54.0 1800芝 良 1:54.8 1-2-3-4 37.0F コリエドール (2.0)	
8	16	ルバ 436kg(-6) 田頭勇貴 佐藤 吉勝(美浦) 父:ファスリエフ 母:ジョウノナンシー(フレンチデピュティ)	370.2 (17人気)	牝2/鹿 54.0kg 勝浦 正樹	16.10.15 東京 未勝利 13 14頭 2番 8人 442kg ▲菊沢一 51.0 1400芝 良 1:25.2 3-4 36.6F レイクキャリー (1.5)	16.09.17 中山 牝未勝利 15 16頭 11番 6人 434kg 和田竜二 54.0 1200ダ 重 1:15.4 7-6 40.7F ブルーヘヴン (3.1)	16.07.24 函館 未勝利 6 13頭 2番 10人 448kg ☆石川裕 53.0 1200芝 良 1:10.8 5-5 35.9F ソレイユフルー (0.9)	16.07.10 函館 未勝利 5 9頭 7番 7人 452kg 吉田隼人 54.0 1200芝 良 1:11.5 5-4 35.7F エスケークラウ (1.1)
8	17	ワシントンレガシー 452kg(+6) 吉田勝己 戸田 博文(美浦) 父:クロフネ 母:フロールデセレッソ(スウェプトオーヴァーボード)	27.5 (8人気)	牝2/芦 54.0kg 柴山 雄一	16.08.21 札幌 新馬 7 11頭 1番 2人 446kg 福永祐一 54.0 1800芝 稍重 1:54.5 8-7-6-7 37.2F ディープウォー (2.4)			
8	18	シルクドレス 506kg(-2) H.H.シェイク・モハメド 黒岩 陽一(美浦) 父:ディープスカイ 母:アークティックシルク(Selkirk)	202.1 (13人気)	牝2/栗 53.0kg ☆石川 裕紀人	16.10.29 新潟 未勝利 11 14頭 10番 4人 508kg △木幡初 52.0 1800ダ 重 2:00.3 5-5-11-9 41.5F ノーブルサター (3.8)	16.10.09 東京 牝未勝利 6 18頭 10番 8人 508kg 内田博幸 54.0 1600芝 重 1:38.5 6-5 36.6F フローレスマジ (1.4)	16.08.14 札幌 新馬 5 7頭 5番 6人 510kg 丸山元気 54.0 1800芝 良 1:54.0 3-3-2-2 36.3F コリエドール (1.2)	

馬柱 (馬毎レース情報)



7	15	クリノミルキー 421.6 (18人気) 444kg(0) 栗本博晴 天間 昭一(美浦) 父:ヨハネスブルグ 母:ビクトワールイゾレ(ダンスインザダーク)	牝2/鹿 54.0kg 江田 照男	16.09.25 未勝利 12 16頭 13番 16人 444kg △原田和 52.0 1:37 6-4-4 アマノカン (0.8)	12 12頭 6番 9人 444kg △原田和 52.0 1200芝 良 1:12.5 7-8 37.2F クセスムーン (2.6)	札幌 6 7頭 6番 7人 450kg 古川吉洋 54.0 1800芝 良 1:54.8 1-2-3-4 37.0F コリエドール (2.0)	
8	16	ピーチメルバ (17人気) 436kg(-6) 田頭勇貴 佐藤 吉勝(美浦) 父:ファスリエフ 母:ジョウノナンシー(フレンチデピュティ)	牝2/鹿 54.0kg 勝浦 正樹	16.10.15 東京 未勝利 13 14頭 2番 8人 442kg ▲菊沢一 51.0 1400芝 良 1:25.2 3-4 36.6F レイクキャリー (1.5)	16.09.17 中山 牝未勝利 15 16頭 11番 6人 434kg 和田竜二 54.0 1200ダ 重 1:15.4 7-6 40.7F ブルーヘヴン (3.1)	16.07.24 函館 未勝利 6 13頭 2番 10人 448kg ☆石川裕 53.0 1200芝 良 1:10.8 5-5 35.9F ソレイユフルー (0.9)	16.07.10 函館 未勝利 5 9頭 7番 7人 452kg 吉田隼人 54.0 1200芝 良 1:11.5 5-4 35.7F エスケークラウ (1.1)
8	17	ワシントンレガシー 27.5 (8人気) 452kg(+6) 吉田勝己 戸田 博文(美浦) 父:クロフネ 母:フロールデセレッソ(スウェプトオーヴァーボード)	牝2/芦 54.0kg 柴山 雄一	16.08.21 札幌 新馬 7 11頭 1番 2人 446kg 福永祐一 54.0 1800芝 稍重 1:54.5 8-7-6-7 37.2F ディープウォー (2.4)			
8	18	シルクドレス 202.1 (13人気) 506kg(-2) H.H.シェイク・モハメド 黒岩 陽一(美浦) 父:ディープスカイ 母:アークティックシルク(Selkirk)	牝2/栗 53.0kg ☆石川 裕紀人	16.10.29 新潟 未勝利 11 14頭 10番 4人 508kg △木幡初 52.0 1800ダ 重 2:00.3 5-5-11-9 41.5F ノーブルサター (3.8)	16.10.09 東京 牝未勝利 6 18頭 10番 8人 508kg 内田博幸 54.0 1600芝 重 1:38.5 6-5 36.6F フローレスマジ (1.4)	16.08.14 札幌 新馬 5 7頭 5番 6人 510kg 丸山元気 54.0 1800芝 良 1:54.0 3-3-2-2 36.3F コリエドール (1.2)	

馬柱 (過去戦績)

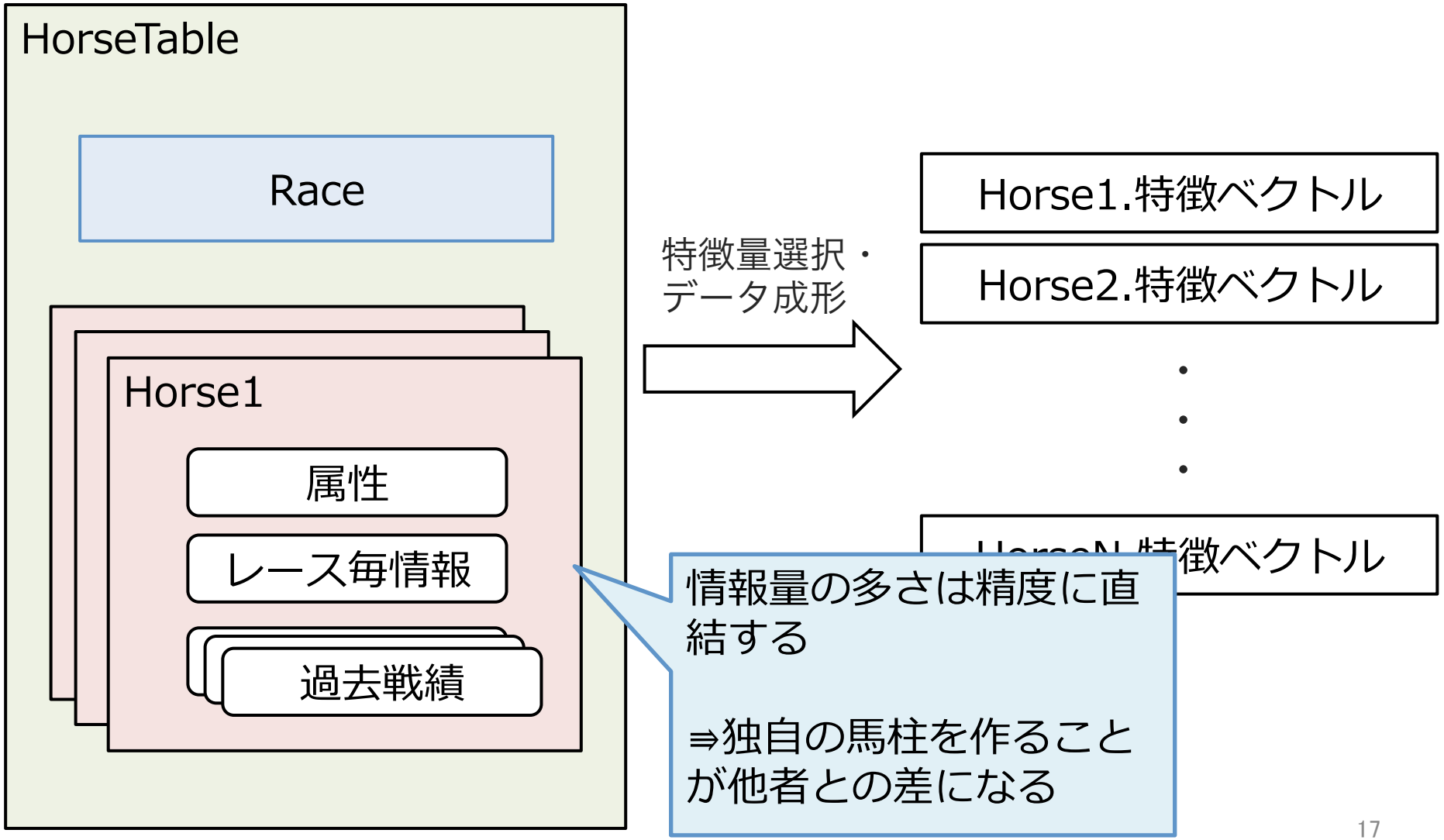


7	15	クリノミルキー 421.6 (18人気) 444kg(0) 栗本博晴 天間 昭一(美浦) 父:ヨハネスブルグ 母:ビクトワールイゾレ(ダンスインザダーク)	牝2/鹿 54.0kg 江田 照男	16.09.25 中山 未勝利 12 16頭 13番 16人 444kg △原田和 52.0 1600芝 稍重 1:37.5 6-4-4 35.9F アマノガワ (0.8)	16.08.20 札幌 未勝利 12 12頭 6番 9人 444kg △原田和 52.0 1200芝 良 1:12.5 7-8 37.2F サクセスムーン (2.6)	16.08.14 札幌 新馬 6 7頭 6番 7人 450kg 古川吉洋 54.0 1800芝 良 1:54.8 1-2-3-4 37.0F コリエドール (2.0)	
		ピーチメルバ		16.10.15 東京 未勝利 13 14頭 2番 8人 442kg ▲菊沢一 51.0 1400芝 良 1:25.2 3-4 36.6F レイクキャリー (1.5)	16.09.17 中山 牝未勝利 15 16頭 11番 6人 434kg 和田竜二 54.0 1200ダ 重 1:15.4 7-6 40.7F ブルーヘヴン (3.1)	16.07.24 函館 未勝利 6 13頭 2番 10人 448kg ☆石川裕 53.0 1200芝 良 1:10.8 5-5 35.9F ソレイユフルー (0.9)	16.07.10 函館 未勝利 5 9頭 7番 7人 452kg 吉田隼人 54.0 1200芝 良 1:11.5 5-4 35.7F エスケークラウ (1.1)
		母:フロールデセレッソ(スウェプトオーヴァーボード)		16.08.21 札幌 新馬 7 11頭 1番 2人 446kg 福永祐一 54.0 1800芝 稍重 1:54.5 8-7-6-7 37.2F ディープウォー (2.4)			
8	18	シルクドレス 202.1 (13人気) 506kg(-2) H.H.シェイク・モハメド 黒岩 陽一(美浦) 父:ディープスカイ 母:アークティックシルク(Selkirk)	牝2/栗 53.0kg ☆石川 裕紀人	16.10.29 新潟 未勝利 11 14頭 10番 4人 508kg △木幡初 52.0 1800ダ 重 2:00.3 5-5-11-9 41.5F ノーブルサター (3.8)	16.10.09 牝未勝利 6 18頭 508kg 内田 54.0 1600芝 重 6-5 36.6F フローレスマジ		

(過去レースごとに)
年月日、コース、芝ダート、距離、条件、頭数、馬番、人気、馬体重、騎手、斤量、馬場状態、着順、タイム、コーナー通過順位、上がり３Fタイム、着差 etc.

過去戦績の長さは馬ごとに異なる
→ 固定長に区切る
→ RNNで時系列データとして扱う

馬柱から特徴ベクトルに



Contents

単純な分類問題

入力:

- レースRにおける出走馬X_iの特徴量 $\mathbf{x}_i \in \Re^m$
- レースRにおける出走馬X_iの成績 $t_i \in \{0,1\}$

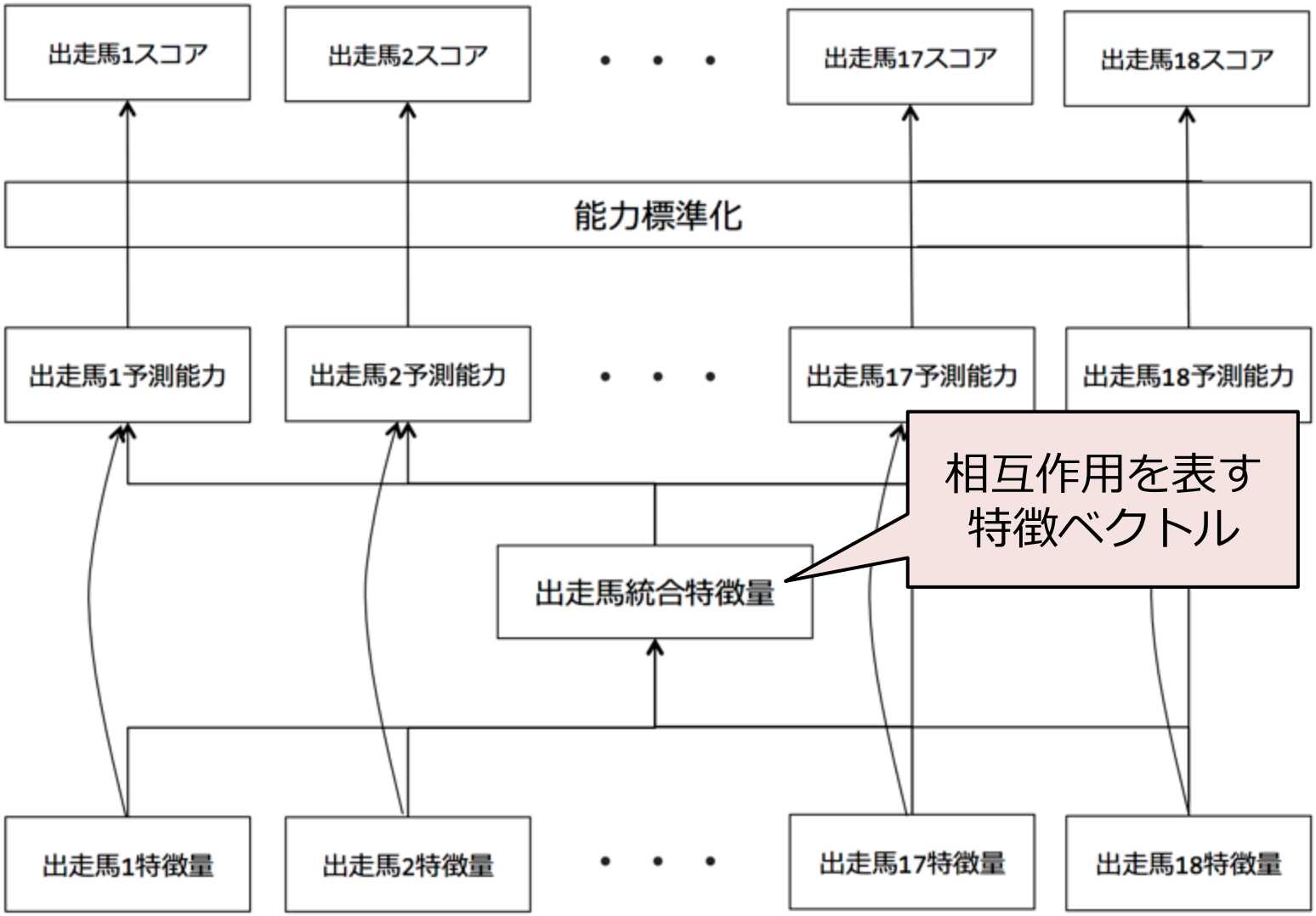
出力:

$t_i = 1$ になる確率 p_i (勝つ確率/スコア)

・1着以内 or not
・着差0.1秒以内 or not

- scikit-learnのモデルで簡単にできる
- 確率が出るので馬券最適化がしやすい
- AlphaKeiba (RandomForest)ではこの問題を解いている
- 出走馬を独立に扱うので相性などの相互作用が考慮されない

相互作用を考慮したモデル



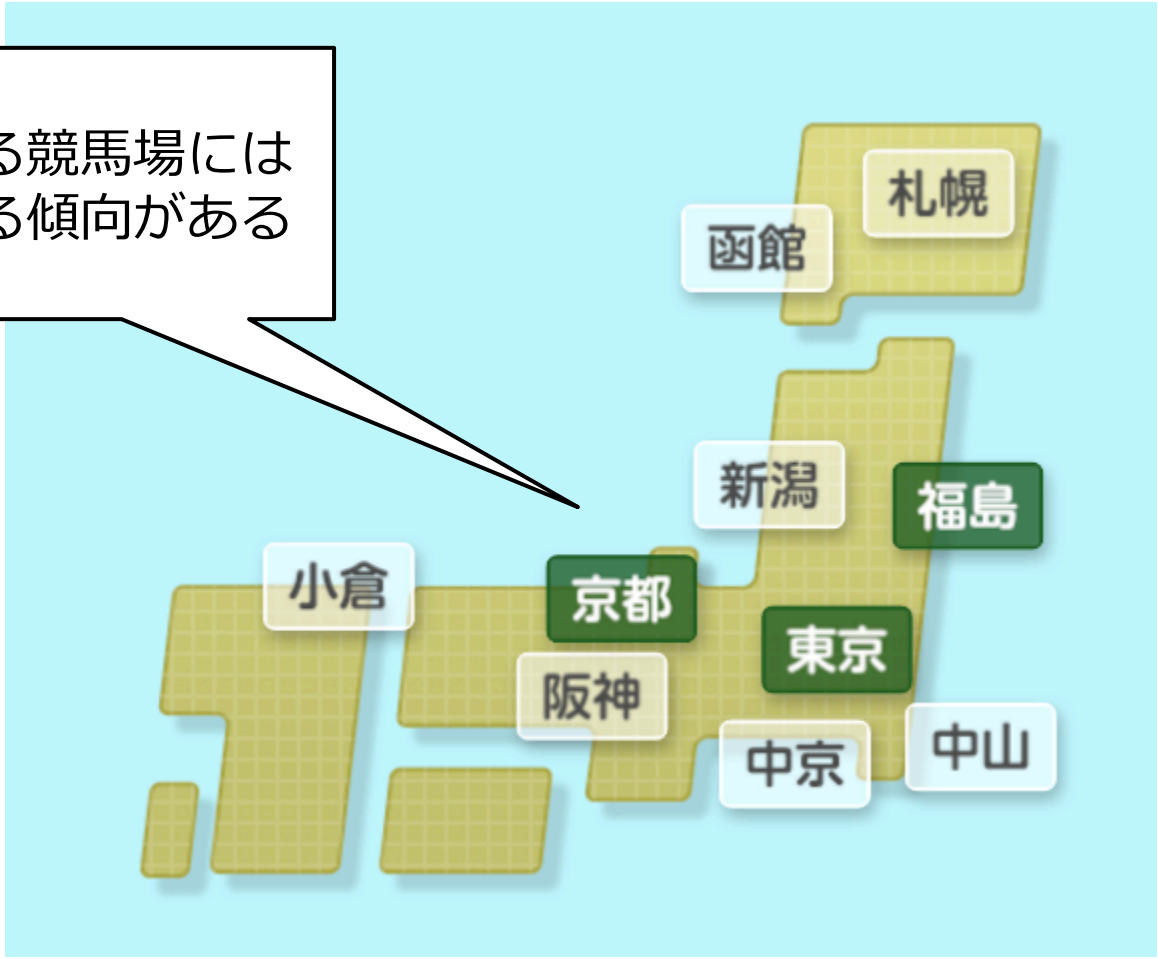
Contents

学習データの分割

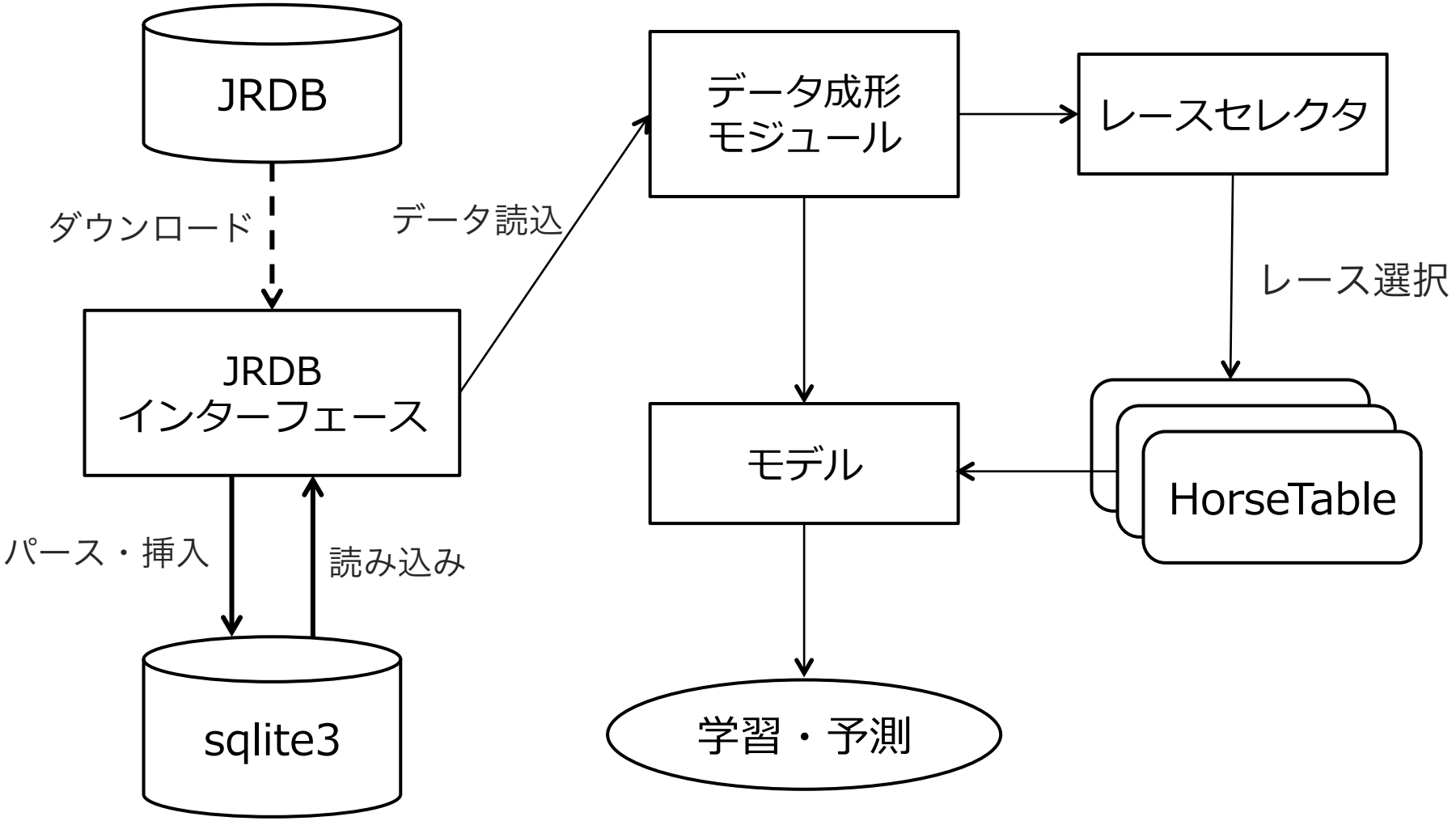
- トラックの種類(芝・ダート・障害)や距離が違えば別の競技
- 特徴量として区別するか学習データを分割するかはノイズと訓練データ量のトレードオフ

競馬場

異なる競馬場には
異なる傾向がある



AlphaImpactの概略図



Contents

1. データの収集
2. 特徴量の作成
3. モデルの設計
4. 学習データの分割
5. 性能評価

的中率・回収率

モデル性能
(東京ダ短距離)

```
---- Top-1 BOX
              hit  hit/true  std (hit/true)      ret  std (all)  std (hit)
win         0.317     0.317           0.465    2.493      9.252     15.091
place       0.707     0.234           0.151    1.193      1.259      1.187

---- Top-2 BOX
                    hit  hit/true  std (hit/true)      ret  std (all)  std (hit)
win               0.585     0.585           0.493    1.877      4.607      5.657
place             0.902     0.402           0.198    1.043      0.728      0.677
quinella place    0.317     0.102           0.152    1.334      2.910      3.823
quinella          0.122     0.122           0.327    2.632     12.128     28.234
exacta            0.122     0.              0.327    3.189     15.848     38.196
```

ベースライン
(単勝人気)

```
---- Top-1 BOX
              hit  hit/true  s                    ll)  std (hit)
win         0.293     0.293                    355      0.962
place       0.805     0.266                    589      0.220

---- Top-2 BOX
                    hit  hit/true  std (hit/true)      ret  std (all)  std (hit)
win               0.512     0.512           0.500    0.852      0.919      0.546
place             0.902     0.413           0.207    0.920      0.466      0.373
quinella place    0.341     0.114           0.158    1.015      1.453      0.610
quinella          0.146     0.146           0.353    1.020      2.595      2.145
exacta            0.146     0.146           0.353    0.959      2.456      2.141
```

複勝・ワイドの正解数に対するリコール

nDCG: Normalized Discounted Cumulative Gain

- ランキング学習の評価でよく使われる指標

$$DCG_k = \sum_{i=1}^{k} \frac{2^{rel_i} - 1}{\log_2 (i+1)}$$

0以上で大きいほど強いことを表すスコア(指数を取る必要はない)
⇛ 着順逆数

下位になるほどスコアの加点が小さくなる

$$nDCG_k = \frac{DCG_k}{IDCG_k}$$

理想的なランキングのときのDCGで割る

まとめ

- データの取得
  - Python使いはJRDBかnetkeibaがおすすめ
- 特徴量の作り方
  - 独自の馬柱を作ることを意識する
- モデルの設計
  - scikit-learnでお手軽に遊ぶなら単純な分類問題
  - さらに精度を目指すなら出走馬の相互作用も入れる
- 学習データの分割
  - ノイズと学習データ量のトレードオフで最良の分割を見つける
- 性能評価
  - 的中率・回収率だけでなくnDCGなど多角的な評価をする