AlphaKeibaはなぜ賞金３０万を獲得できたのか？

東工大ブライアンズトウコウ
リーダーのぬくい
twitter: @heartz2001

自己紹介

名前: 貫井 駿

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

サークル: 東工大競馬研究会 部長

専門: ネットワーク解析・機械学習

メンバー紹介

チーム名: ブライアンズトウコウ

中条	貫井	小峰
(競馬マニア)	(データ解析担当)	(エンジニア)

今日の内容

- 電脳賞について
- 『AlphaKeiba』の構成
- データ成形
- 予想ロジック
- 馬券最適化
- デモ

電脳賞(春)について

- (株)ドワンゴ主催競馬予想アルゴリズム大会
- 回収率の部と1着予想の部で競う
- 通算回収率100%越えると賞金**30万円**
- 50R/50R的中させると20万円(?!)

電脳賞ルール

- 1ヶ月間計50レースが対象
- 毎開催日5レース選出
- ソフトのみを使ってレース結果を予想
- 12,000pt/Rを使い切る
- JRA-VANが提供するData Lab.のデータを使用する

電脳賞成績

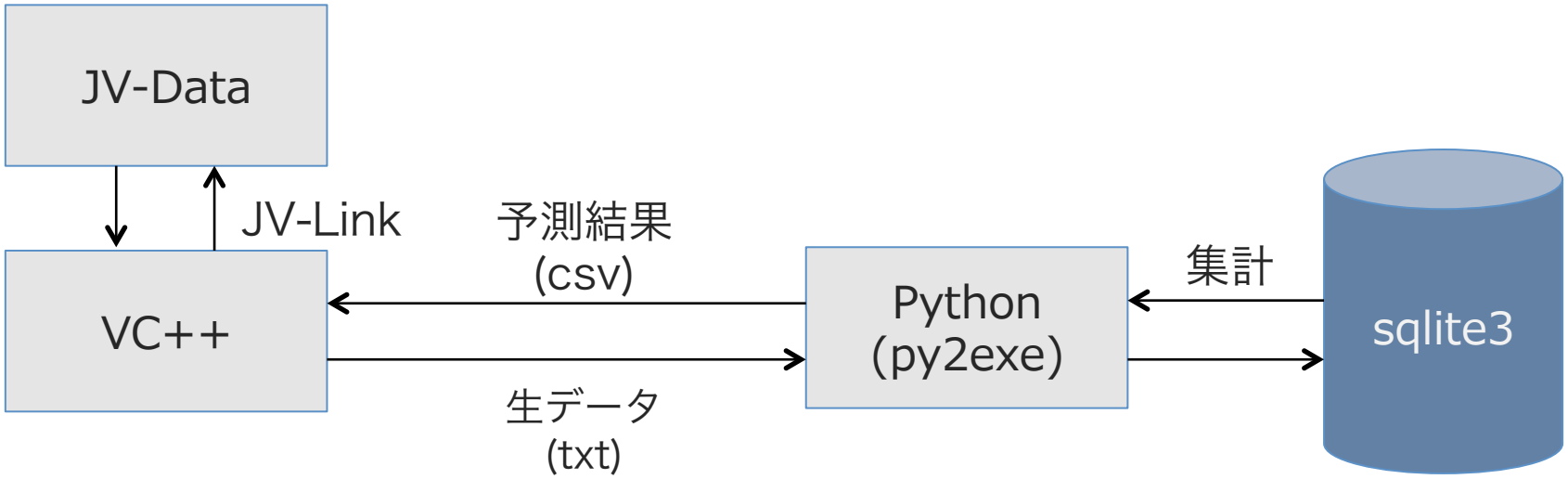
	東工大チーム		お茶女チーム	
	1着予想の部	回収率の部	1着予想の部	回収率の部
通算	14R/50R	100.6%	18R/50R	90.6%
3月5日	1R/5R	70.9%	2R/5R	188.0%
3月6日	0R/5R	54.6%	0R/5R	0%
3月12日	3R/5R	132.3%	2R/5R	23.0%
3月13日	2R/5R	256.7%	1R/5R	98.0%
3月19日	2R/5R	132.2%	2R/5R	84.7%
3月20日	1R/5R	0%	3R/5R	22.9%
3月21日	1R/5R	0%	0R/5R	78.7%
3月26日	1R/5R	91.0%	4R/5R	266.2%
3月27日	2R/5R	198.5%	2R/5R	22.3%
4月10日	1R/5R	70.1%	2R/5R	120.0%

30万への道のり

AlphaKeibaとは

- 機械学習アルゴリズムによる馬券予想システム
- JRA-VAN公式データを利用
- α指数の計算と馬券最適化ができる

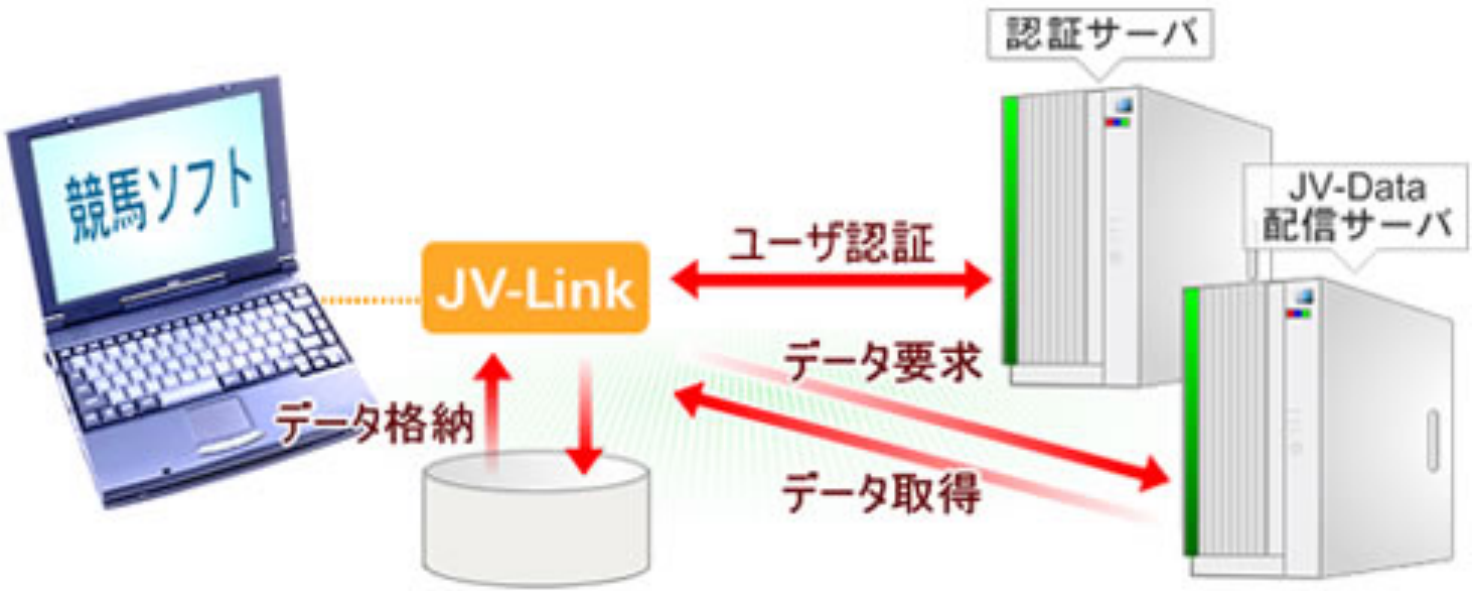
AlphaKeibaの構成



データ成形

データ取得

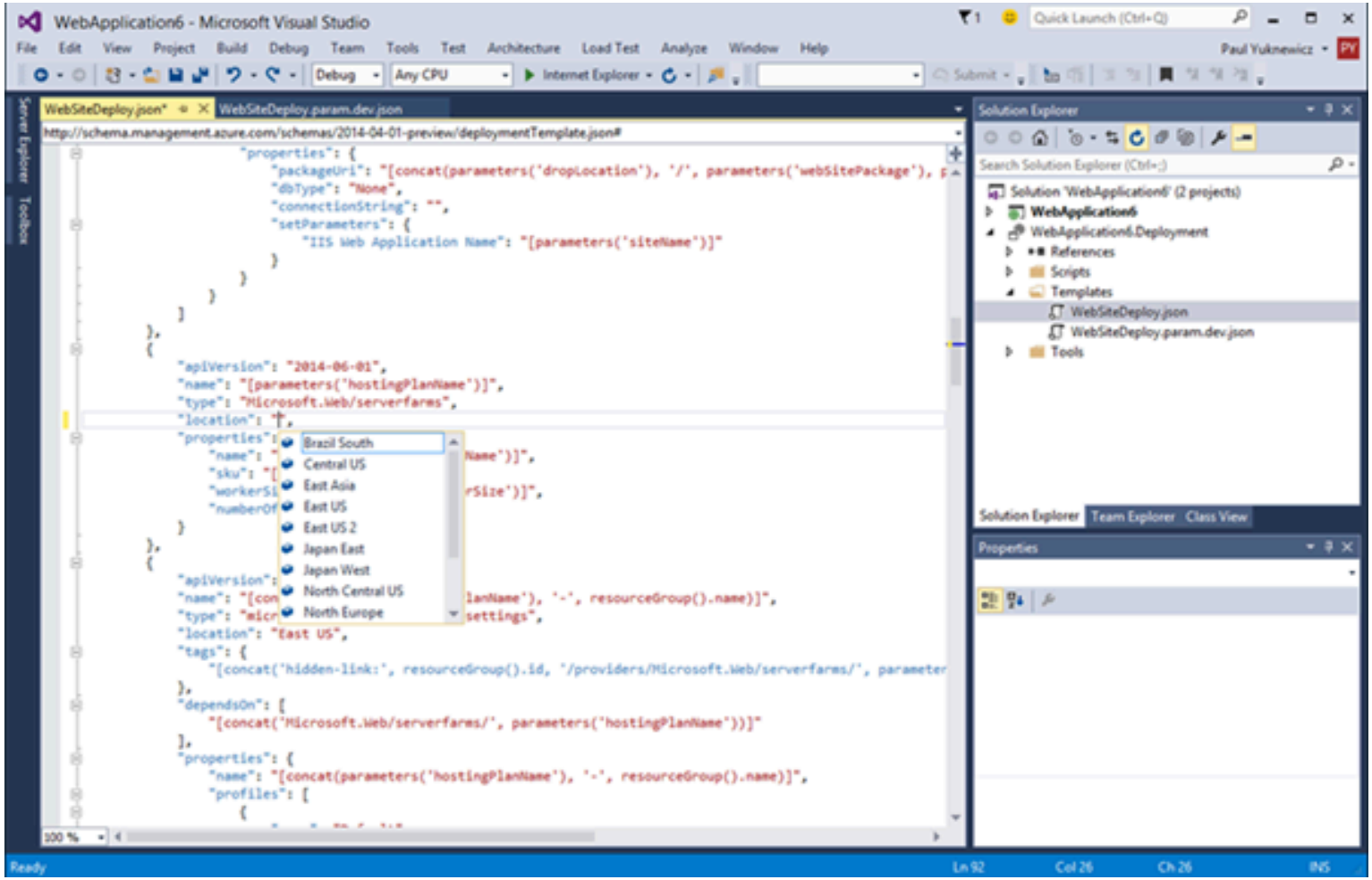
- JV-Link APIを利用
- 月額2,052円(税込)で利用可
- データを使える形にするためにいろんな困難が待ち受けている



※JRA-VAN HPより

困難その１

- Visual C++ (or C#)を使わなくてはならない(マニュアルがVS 2005)



困難その2

・固定長テキスト形式で提供されるデータ

```
SE7201606062016060405030101101201310199 0クリノリーズ
0021010310103 1伊藤伸一970006栗本　博晴
緑，黒袖
5200000001151000 00井上敏樹
20422+0120131300128313 4
0000040625861500000000000000000000000390201310554 6
ミッキーオフィサー                              0000000000
0000000000
+0240312951002400071 53
SE7201606062016060405030101102201310486 7プラウドフェロー
0021010310105 2菊川正達013803ディアレストクラブ
薄紫，青一本輪，袖青縦縞
5400000000689000 0江田照男          00424-048014140012862
0000161517171300000000000000000000003822013105546
ミッキーオフィサー                              0000000000
0000000000
+0270312859002400071 24
…
```

データのパース

・定義書(excel)をパースしてそれを基に生データをパースする
(後にデータをパースするSDKが提供されていることを知りました)

3．馬毎レース情報				レコード長	555 バイト				項番2. データ区分									
項番		キー	項目名	位置	繰返	バイト	合計	初期値	説明	1	2	3	4	5	6	7	A	B1	B2
1			レコード種別ID	1		2			"SE" をセットレコードフォーマットを特定する	○	○	○	○	○	○	○	○	○	○
2			データ区分	3		1		0	1:出走馬名表(木曜) 2:出馬表(金・土曜) 3:速報成績(3着まで確定) 4:速報成績(5着まで確定) 5:速報成績(全馬着順確定) 6:速報成績(全馬着順+コーナ通過順) 7:成績(月曜) A:地方競馬 B:海外国際レース 9:レース中止 0:該当レコード削除(提供ミスなどの理由による)	○	○	○	○	○	○	○	○	○	○
3			データ作成年月日	4		8		0	西暦4桁＋月日各2桁 yyyymmdd 形式	○	○	○	○	○	○	○	○	○	○
4		○	開催年	12		4		0	該当レース施行年 西暦4桁 yyyy形式	○	○	○	○	○	○	○	○	○	○
5		○	開催月日	16		4		0	該当レース施行月日 各2桁 mmdd形式	○	○	○	○	○	○	○	○	○	○
6		○	競馬場コード	20		2		0	該当レース施行競馬場 <コード表 2001.競馬場コード>参照	○	○	○	○	○	○	○	○	○	○
7		○	開催回[第N回]	22		2		0	該当レース施行回 その競馬場でその年の何回目の開催かを示す	○	○	○	○	○	○	○	○	-	-
8		○	開催日目[N日目]	24		2		0	該当レース施行日目 そのレース施行回で何日目の開催かを示す	○	○	○	○	○	○	○	○	-	-
9		○	レース番号	26		2		0	該当レース番号	○	○	○	○	○	○	○	○	△	△
10			枠番	28		1		0		-	○	○	○	○	○	○	○	-	-
11		●	馬番	29		2		0	出走馬名表時点では、全て初期値を設定（出馬表時点で馬番が確まるため、馬番を純粋にキー設定していると、同一馬の情報が重複する） しかし、海外国際レース等、出走メンバーの血統登録番号が初期値の場合があるため、馬番をキーとして使用する。(海外国際レースなどで馬番情報がない場合は任意に連番を設定する)	-	○	○	○	○	○	○	○	△	△
12		○	血統登録番号	31		10		0	生年(西暦)4桁＋品種1桁<コード表2201.品種コード>参照＋数字5桁	○	○	○	○	○	○	○	○	○	-
13			馬名	41		36		S sp	通常全角18文字。海外レースにおける外国馬の場合のみ全角と半角が混在	○	○	○	○	○	○	○	○	○	○
14			馬記号コード	77		2		0	<コード表 2204.馬記号コード>参照	○	○	○	○	○	○	○	-	○	-
15			性別コード	79		1		0	<コード表 2202.性別コード>参照	○	○	○	○	○	○	○	○	○	△
16			品種コード	80		1		0	<コード表 2201.品種コード>参照	○	○	○	○	○	○	○	○	○	-
17			毛色コード	81		2		0	<コード表 2203.毛色コード>参照	○	○	○	○	○	○	○	○	○	-
18			馬齢	83		2		0	出走当時の馬齢 （注）2000年以前は数え年表記 2001年以降は満年齢表記	○	○	○	○	○	○	○	○	○	△
19			東西所属コード	85		1		0	<コード表 2301.東西所属コード>参照	○	○	○	○	○	○	○	-	○	-
20			調教師コード	86		5		0	調教師マスタへリンク	○	○	○	○	○	○	○	-	○	-
21			調教師名略称	91		8		S	全角4文字	○	○	○	○	○	○	○	○	○	-
22			馬主コード	99		6		0	馬主マスタへリンク	○	○	○	○	○	○	○	-	○	-
23			馬主名(法人格無)	105		64		S sp	全角32文字 ～ 半角64文字 （全角と半角が混在） 株式会社、有限会社などの法人格を示す文字列が頭もしくは末尾にある場合にそれを削除したものを設定。また、外国馬主の場合は、馬主マスタの8.馬主名欧字の頭64バイトを設定	○	○	○	○	○	○	○	△	○	-
24			服色標示	169		60		S	全角30文字 馬主毎に指定される騎手の勝負服の色・模様を示す (レーシングプログラムに記載されているもの) (例)"水色、赤山形一本輪、水色袖"	○	○	○	○	○	○	○	-	○	-
25			予備	229		60		S		○	○	○	○	○	○	○	○	○	○
26			負担重量	289		3		0	単位0.1kg	○	○	○	○	○	○	○	○	○	-
27			変更前負担重量	292		3		0	なんらかの理由により変更された場合のみ変更前の値を設定	-	-	○	○	○	○	○	-	-	-
28			ブリンカー使用区分	295		1		0	0:未使用 1:使用	○	○	○	○	○	○	○	-	-	-
29			予備	296		1		0		○	○	○	○	○	○	○	○	○	○
30			騎手コード	297		5		0	騎手マスタへリンク	○	○	○	○	○	○	○	-	△	-
31			変更前騎手コード	302		5		0	なんらかの理由により変更された場合のみ変更前の値を設定	-	-	○	○	○	○	○	-	-	-
32			騎手名略称	307		8		S	全角4文字	○	○	○	○	○	○	○	○	○	-
33			変更前騎手名略称	315		8		S	なんらかの理由により変更された場合のみ変更前の値を設定	-	-	○	○	○	○	○	-	-	-
34			騎手見習コード	323		1		0	<コード表 2303.騎手見習コード>参照	○	○	○	○	○	○	○	-	-	-

Python+Pandasで成形

	RaceID	GateNo	HorseID	SexCD	ColorCD	Age	WestEast	TrainerCD	OwnerCD	Burden	...	TurfConditionCD	DirtConditio
0	2015010406010101	1	2012101800	2	4	3	1	1076	829800	540	...	0	1
1	2015010406010101	2	2012101349	2	1	3	1	1100	936800	540	...	0	1
2	2015010406010101	3	2012103557	1	3	3	1	420	639009	560	...	0	1
3	2015010406010101	4	2012100085	1	4	3	1	1063	149002	560	...	0	1
4	2015010406010101	5	2012105911	1	1	3	1	1127	849030	560	...	0	1
5	2015010406010101	6	2012103681	2	3	3	1	1103	629800	540	...	0	1
6	2015010406010101	7	2012100472	1	7	3	1	1007	476800	560	...	0	1
7	2015010406010101	8	2012100783	1	3	3	1	1147	573033	560	...	0	1
8	2015010406010101	9	2012104273	2	5	3	1	1097	19033	540	...	0	1
9	2015010406010101	10	2012103776	2	3	3	1	1089	748009	540	...	0	1
10	2015010406010101	11	2012101002	2	3	3	1	1026	384033	540	...	0	1
11	2015010406010101	12	2012101100	1	4	3	1	1005	270006	560	...	0	1
12	2015010406010101	13	2012100821	2	1	3	1	1080	547800	540	...	0	1
13	2015010406010101	14	2012100011	1	1	3	1	1093	418030	560	...	0	1
14	2015010406010101	15	2012105529	2	7	3	1	1010	313008	510	...	0	1
15	2015010406010101	16	2012101183	1	1	3	1	392	540033	560	...	0	1
16	2015010406010102	1	2012101132	1	1	3	1	1147	901007	560	...	0	1
17	2015010406010102	2	2012100539	1	3	3	1	1133	338009	530	...	0	1
18	2015010406010102	3	2012103940	1	3	3	1	1094	755030	560	...	0	1
19	2015010406010102	4	2012109142	1	3	3	1	435	872006	560	...	0	1
20	2015010406010102	5	2012102302	1	3	3	1	1052	170800	560	...	0	1
21	2015010406010102	6	2012105765	1	3	3	1	1131	674004	560	...	0	1
22	2015010406010102	7	2012103103	1	3	3	1	423	966007	560	...	0	1

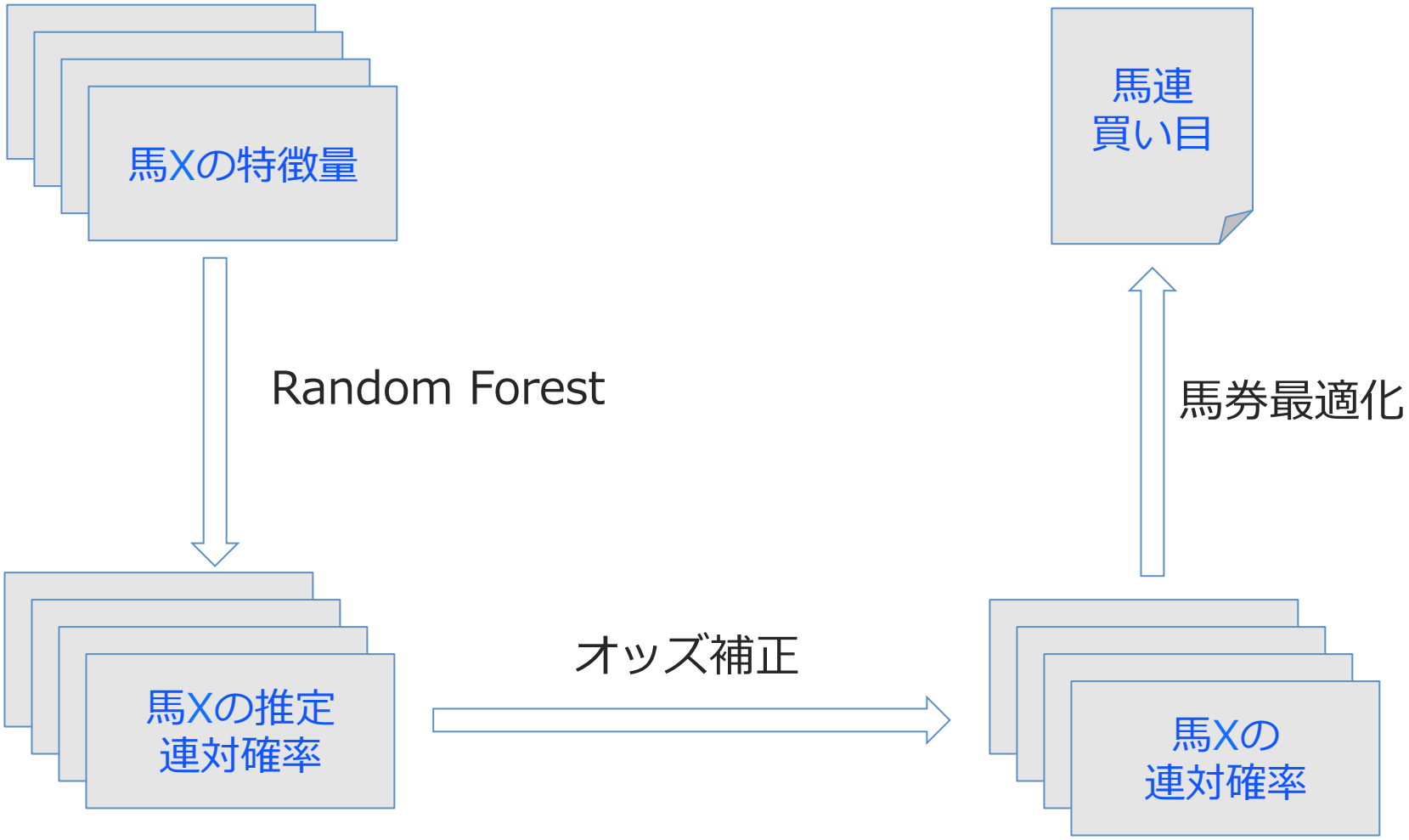
Data Lab.の提供データ

提供されているデータの例

- 1990年以降の中央レース成績
  - ハロンタイム,着順,着差,脚質 etc
- オッズ
  - 最終オッズ,速報オッズ,時系列オッズ
- 血統
- 速報データ(馬場状態, 天気など)
- 調教タイム

予想ロジック

ロジックの流れ



主な特徴量

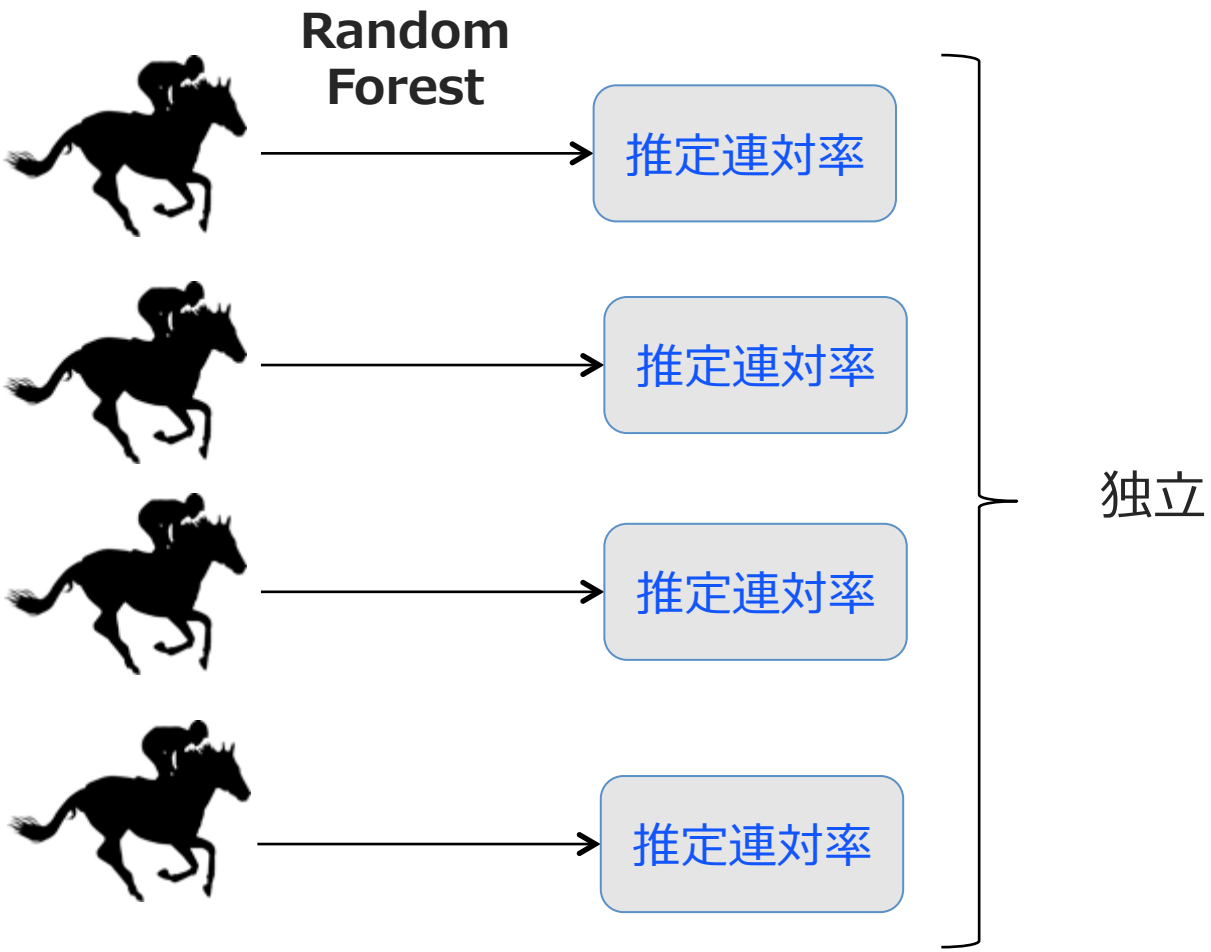
- 出走頭数
- 当該馬の出走回数
- 過去3レースの着順
- 過去3レースの人気
- 過去3レースの着差
- 過去3レースの脚質
- 過去3レースの距離
- 過去3レースのクラス
- 出走馬の平均脚質
- 調教タイム
- 騎手の平均着順
- 騎手の平均脚質

Random Forest

- 出走頭数
- 当該馬の出走回数
- 過去3レースの着順
- 過去3レースの人気
- 過去3レースの着差
- 過去3レースの脚質
- 過去3レースの距離
- 過去3レースのクラス
- 出走馬の平均脚質
- 調教タイム
- 騎手の平均着順
- 騎手の平均脚質

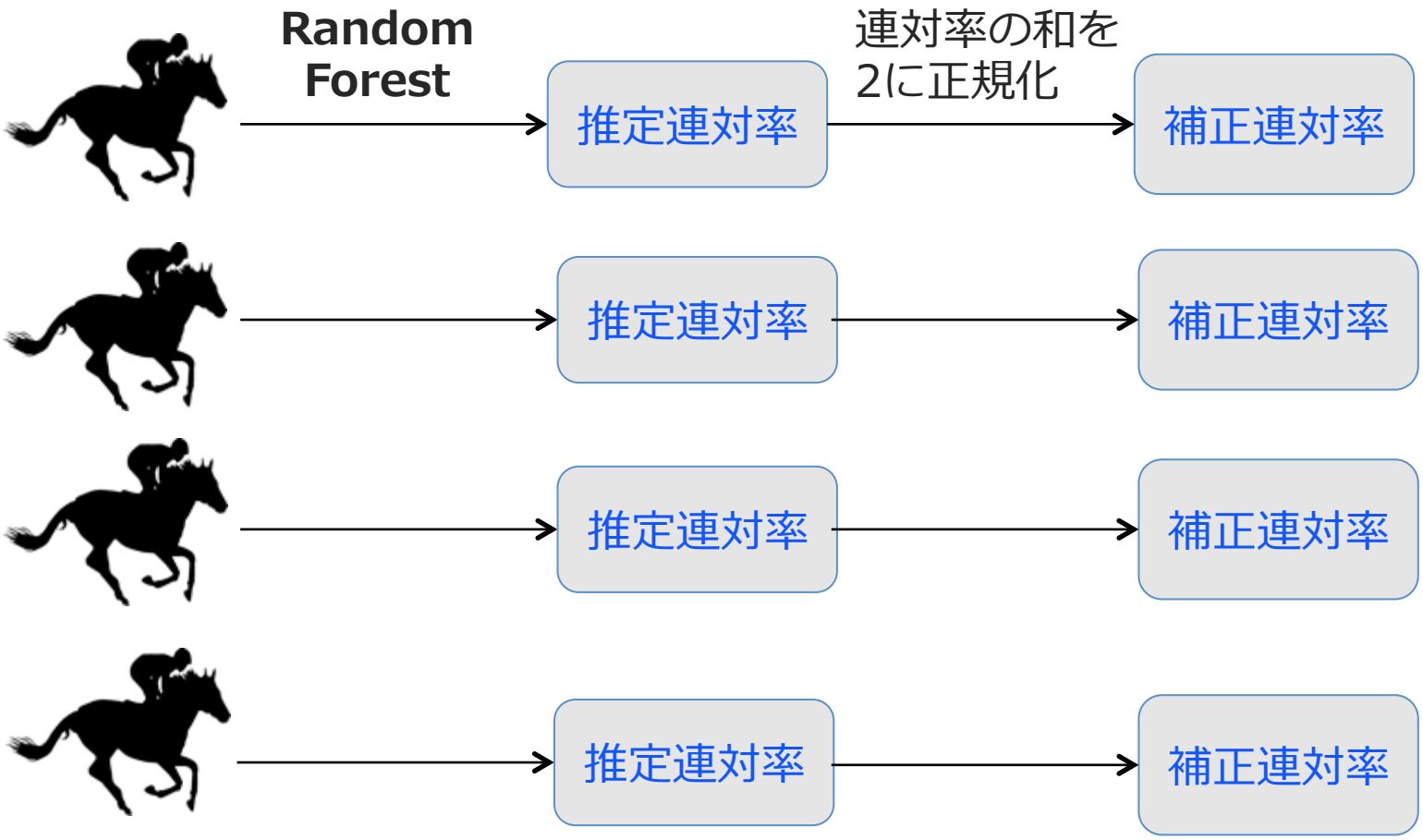
連対率の推定

- 馬連に最適化するために連対率を推定



連対率の推定

- 馬連に最適化するために連対率を推定



オッズによる補正

- 馬連の周辺支持率から連対確率を計算

$$\text{馬連(X-Y)支持率} = \frac{0.748}{\text{馬連(X-Y)オッズ}}$$

控除率

$$\text{馬X連対確率} = \sum_{Y} \text{馬連(X-Y)支持率}$$

馬券最適化

馬券最適化アルゴリズム

入力: 推定連対率{p_X}
馬連オッズ{o_{XY}}
最低的中率 p_{min}
予算 M (=12,000pt)

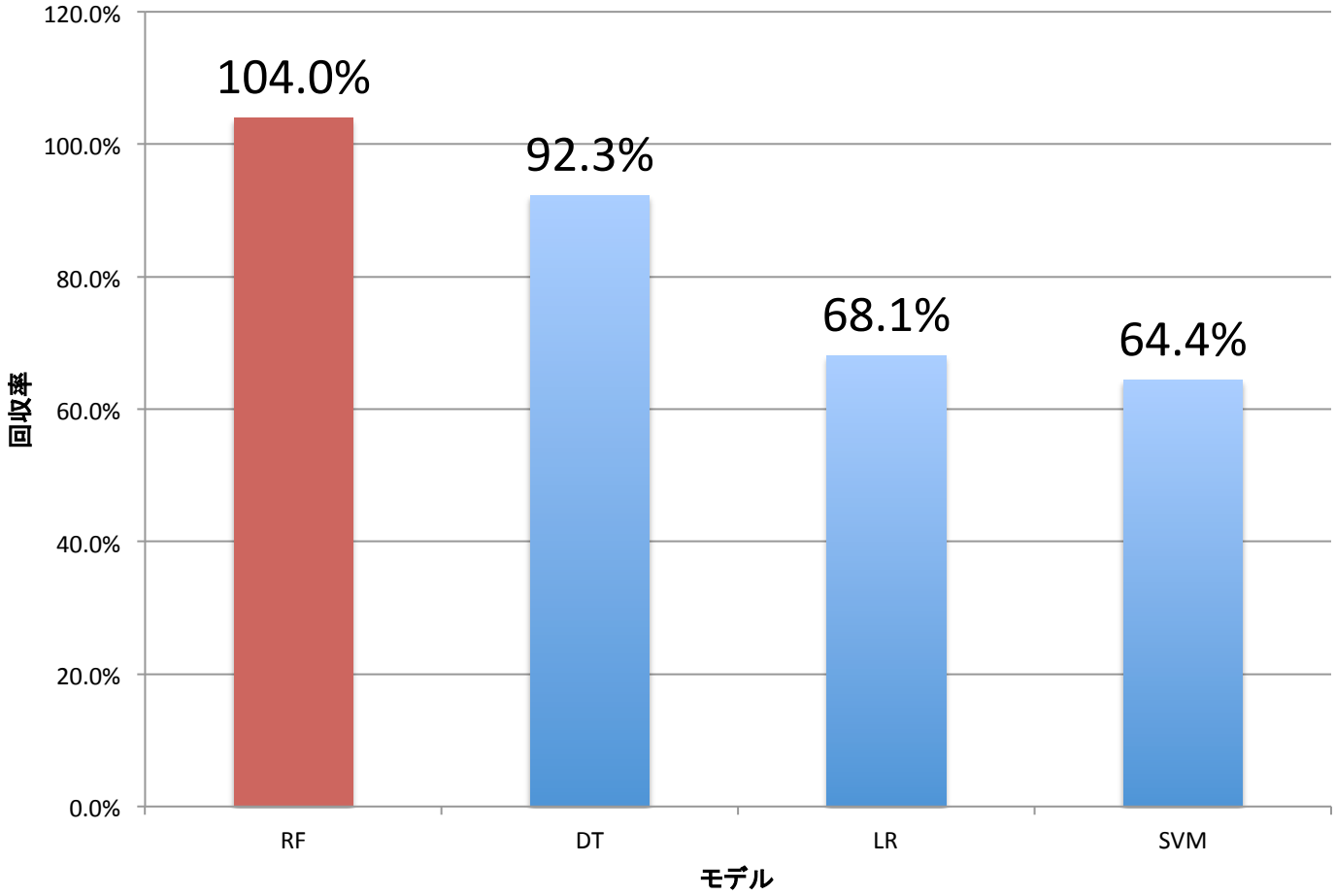
出力: 期待値を最大化する買い目(金額) B={(b_{XY},m_{XY})}

試したモデル

- ロジスティック回帰
- SVM
- 決定木
- Random Forest

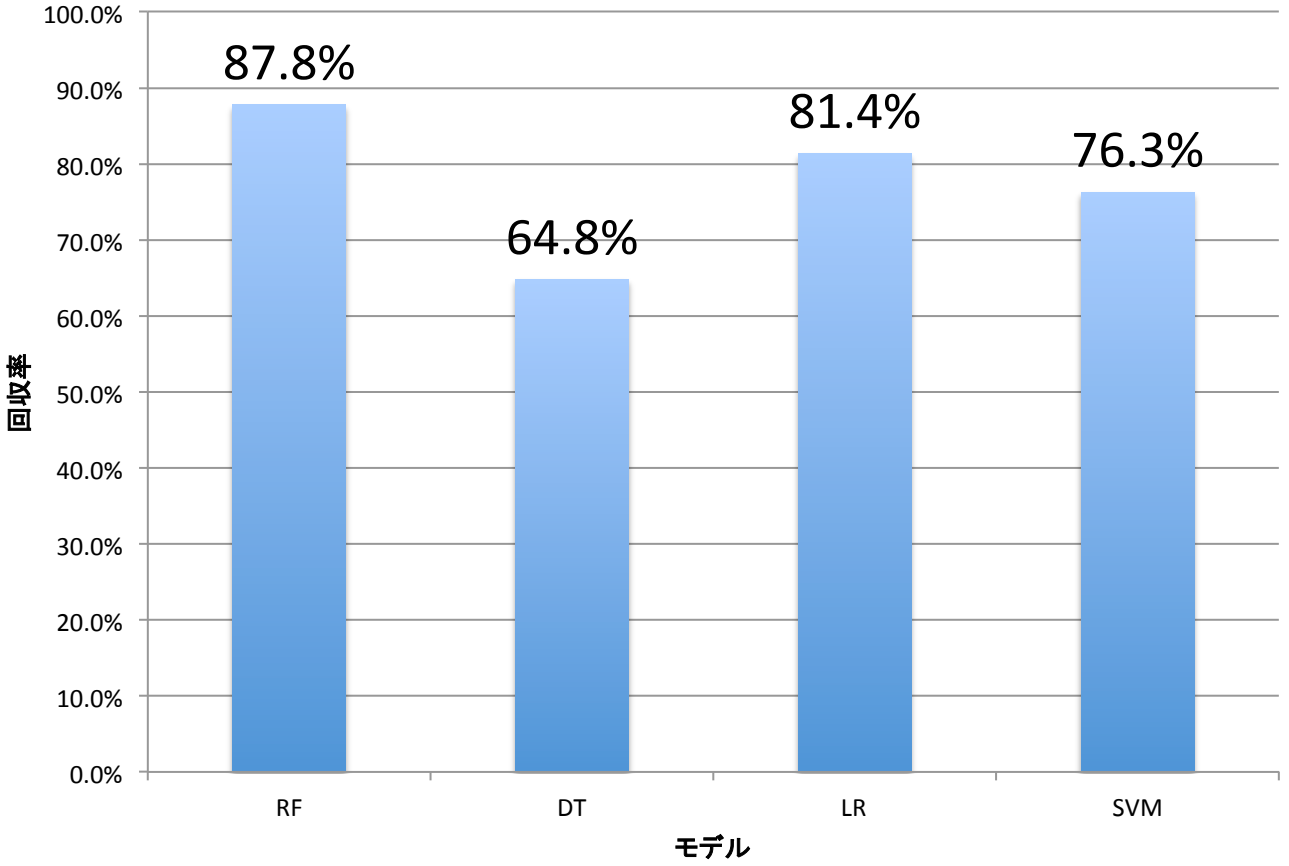
シミュレーション結果

- 阪神ダート1800m(1050レース)
- 訓練サンプル 約400R/回



シミュレーション結果

- 中距離ダート(1008レース)
- 訓練サンプル 約4500R/回



デモ

AlphaKeiba on Azure

まとめ

- データ成形は大変
- 今のところRandomForestが1番強い
- RandomForest+馬券最適化と組み合わせたら回収率100%超えた

今後の展望

- 馬券最適化を強化学習でやる
- 流行りのディープラーニングを使いたい
- 可視化ツール作りたい

さいごに


AlphaKeibaの
穴馬マイニング
電脳賞
公認
競馬データ解析プログラムがはじき出した
注目馬を毎週お届けします！

ご清聴ありがとうございました