

# 競馬でわかる AI リテラシー入門

— Ask・Check・Decide で“未来”を読む 30 分 —

株式会社AlphaImpact

大元 司



# 自己紹介：AIで“勝ち筋”を読む人

- 氏名：大元 司（おおもと つかさ）
- 所属：株式会社AlphaImpact
- 役職：取締役 & 共同創業者
- 経歴ハイライト
  - 株式会社ドワンゴ(2015-2019)
    - 動画 / 配信系サービスで ML 活用
    - オンライン学習サービス「N予備校」教材開発
  - 株式会社AlphaImpactにて、競馬予測人工知能の開発と運用（2019-現在）

# 本日の講演内容

- ○ 話すこと
  - 競馬を題材に体験する AIリテラシー 3 ステップ
    - Ask — ゴールと問いを具体化する
    - Check — データを疑い、可視化・検証する
    - Decide — 確率とリスクを踏まえ意思決定する
  - AIと人間の役割分担、モデルの限界と責任
- ✕ 話さないこと
  - 的中率を上げる“競馬 AI”の実装レシピ
  - 馬券必勝法・儲け方の指南
    - ※ 馬券購入は 20 歳から！

競馬ってそもそも何？



JRA公式YouTubeチャンネルより

[https://www.youtube.com/watch?v=fEqV1sn\\_tE8](https://www.youtube.com/watch?v=fEqV1sn_tE8)

# 競馬 = “2分間のデータ競技”

- 超短時間決着
  - 1 レースは約 2、3分 → 瞬時に勝敗が確定
  - 1 開催日で 24~36レース
- 豊富なデータが毎週更新
  - 過去成績・ラップタイム・オッズ・馬体重・天候 …
  - 週末ごとに 数千件 の新データが追加 → 機械学習と相性◎

2025年6月1日 (日曜) 2回東京12日 発走時刻: 15時40分 天候: 晴 芝: 良

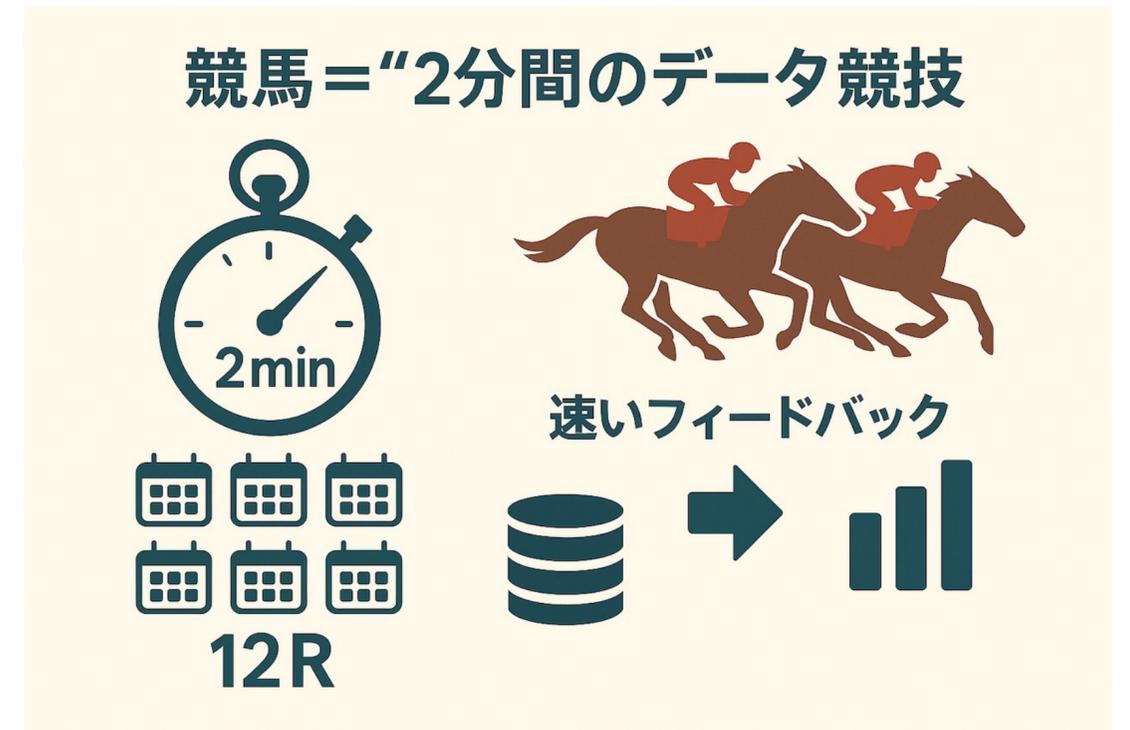
**11R** 第92回 東京優駿 **GI**  
**WIN5** 3歳 オープン (国際) 牡・牝 (指定) 馬齢 コース: 2,400メートル (芝・左)

本賞金 (万円) 1着 30,000 2着 12,000 3着 7,500 4着 4,500 5着 3,000  
 付加賞 (万円) 1着 2,771.3 2着 791.8 3着 395.9

枠	馬番	馬名 / 単勝オッズ(人気) 戦績 / 総賞金 / 馬体重 馬主名 / 調教師名 / 血統	性齢/毛色 負担重量 騎手名 プレレーティング	前走	前々走	3
1	1	<b>リラエンブレム</b> (2.0.0.1) 4879.5万 <b>484kg (+4)</b> (株)Gリビエール・レーシング 武幸四郎 (栗東) 父: キズナ 母: デルフィニア2 (母の父: Galileo)	 牡3/黒鹿 57.0kg 浜中俊 111 M	2025年3月29日 阪神 <b>毎日杯</b> 7着 浜中俊 1800芝 良 103 5 5 ファンダム (0.8)	2025年1月13日 中京 <b>シンザン記念</b> 1着 浜中俊 1600芝 良 111 4 8 9 アルテヴェローチェ (0.4)	2024年10月 <b>新馬</b> 1着 浜中俊 1600芝 良 5 3 スマートミス
1	2	<b>ショウヘイ</b> (2.1.0.1) 6934.2万 <b>460kg (+4)</b> 石川達絵 友道康夫 (栗東) 父: サートゥルナーリア 母: オーロトラジェ (母の父: オルフェーヴル)	 牡3/黒鹿 57.0kg C.ルメール 112 L	2025年5月10日 京都 <b>京都新聞杯</b> 1着 川田将雅 2200芝 稍重 112 2 2 2 2 エムズ (0.4)	2025年2月9日 京都 <b>きさらぎ賞</b> 4着 川田将雅 1800芝 稍重 101 5 4 サトノシャイニング (1.0)	2024年12月 <b>未勝利</b> 1着 川田将雅 1800芝 良 3 4 フィーユチャ
2	3	<b>エリキング</b> (3.0.0.1) 5662万 <b>500kg (0)</b> 藤田晋 中内田充正 (栗東) 父: キズナ 母: ヤングスター (母の父: High Chaparral)	 牡3/鹿 57.0kg 川田将雅 112	2025年4月20日 中山 <b>皐月賞</b> 11着 川田将雅 2000芝 良 107 9 8 11 10 ミュージアムマイル (1.0)	2024年11月23日 京都 <b>京都2歳S</b> 1着 川田将雅 2000芝 良 112 4 5 3 3 ジョバンニ (0.2)	2024年9月 <b>野路菊S</b> 1着 川田将雅 2000芝 良 3 3 3 ジョバンニ

# 競馬 = “2分間のデータ競技”

- 参加者 = “データで戦う” プレーヤー
  - 馬・騎手・調教師: パフォーマンスを数値で管理
  - 観客: 馬券を買う 予測投資家 (100 円から参戦可)
- 巨大市場規模
  - 例: 有馬記念の馬券売上 約 550 億円
  - 売上はパリミュチュエル方式での的中者へ分配 → 確率とお金が直結
- AI 活用の魅力
  - フィードバックが速い: 2分後に答え合わせ
  - データ量 × 回転数 が多い → 素早い学習・検証サイクルが可能



# “儲け方”を決める = ゴールを決める

- 馬券の種類ごとに“問い”が変わる
  - 単勝：1着を当てる(的中率重視・配当小)
  - 馬単／馬連：1-2着を順・不順で当てる(中リスク中リターン)
  - 三連単：1-3着順を当てる(的中率低・配当大)
- ゴール設定のバリエーション
  - 「月末残高+1万円」→ 回収率 & 資金管理が最優先
  - 「的中率でクラス1位」→ 単勝・複勝で精度追求
  - 「万馬券でSNS映え」→ 三連単でハイリスク志向

券種	的中率の目安	平均配当の目安	戦略キーワード
単勝	30-35 %	3-5倍	安定・薄利多売
馬単	2-3%	30-45倍	中間的
三連単	0.5-1.0%	700-1000倍	一撃必殺

“どんな儲け方をしたいか” が “良い問い” の出発点

# Ask — 良い問いを作る

当たるって何？

# 当てるって何？

- 「当てる」って具体的にどういうこと？
  - 仮に「1着馬を当てること」とする
    - 「1着」って何？
      - 1位の馬 → 2位以下の馬が存在する
        - 1着かどうかの分類問題！
        - (例)18頭から1頭を選ぶ
      - 1番早くゴールを駆け抜けた馬
        - 走破タイムを予測する回帰問題！
        - (例)1分28秒3を予測する
- 漠然とした「当てる」にも様々な解釈と問題定義が考えられる



# 当てるって何？ — ターゲット定義

- ラベル予測(分類)
  - 例:「1着になる馬はどれ？」
  - 出力:1着/その他の2クラス
- 値予測(回帰)
  - 例:「各馬の走破タイムは？」
  - 出力:実数タイム(秒)
- 順位予測(ランキング)
  - 例:「タイム順に並べたトップ3は？」
  - 出力:順位リスト(n頭)



# 評価指標を合わせる

- ターゲット別に適切な指標を選ぶ
  - ラベル予測 → Accuracy / LogLoss
  - 値予測 → RMSE / MAE
  - 順位予測 → NDCG / MRR
- 指標ミスマッチの落とし穴
  - 的中率90 %でも回収率50 %なら赤字
  - RMSEが小さくても上位3頭の順序が全滅することも
- 測定の一貫性が肝心
  - 期間・サンプル数・クロスバリデーション条件をそろえる
  - “学習セットと同じ指標で本番を評価”が基本

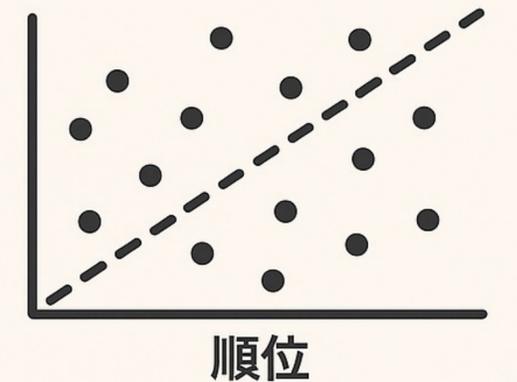
## Accuracy

正解率が高いほど良い



## RMSE

二乗誤差が小さいほど良い



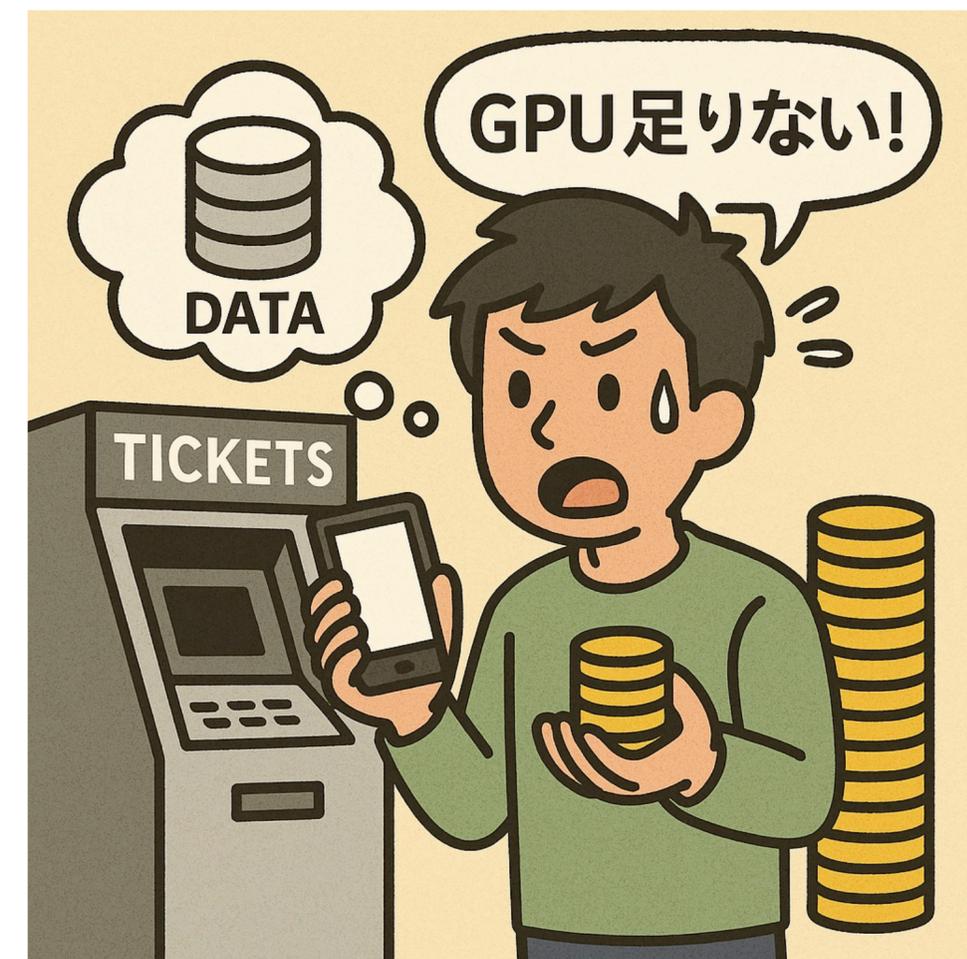
## NDCG

上位ほど正解に近いほど良い



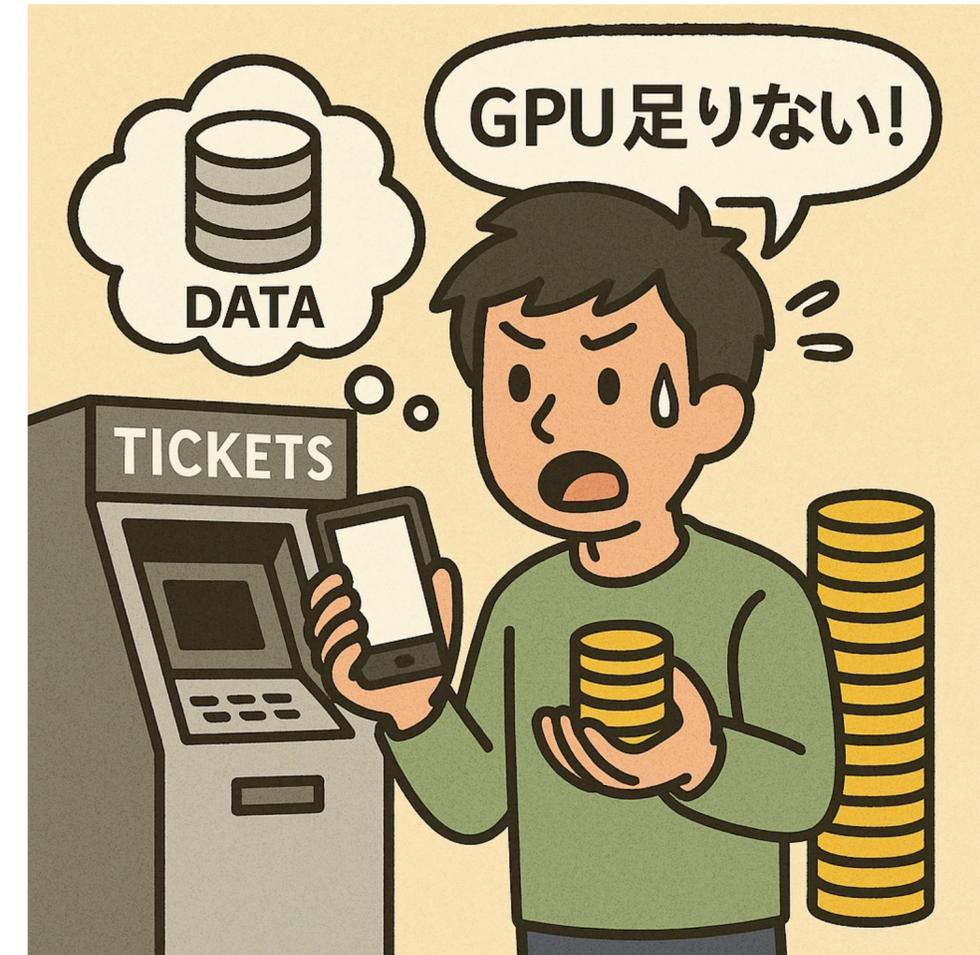
# 制約と現実を忘れない

- データ制約
  - 発走前に取得できる情報だけ(調教タイム・当日馬体重など)
  - “最終オッズ”やレース結果は学習時のみ、本番では見えない
- 資金制約
  - 例:学生なら 1 日 1,000 円以内 / 月 1 万円以内など上限を設定
  - ケリー基準などの資金管理ルールも事前に決めておく
- 計算リソース制約
  - スマホ単体で 1 分以内に推論できる軽量モデルが条件
  - 通信環境(4G / オフライン)でも動く必要があるか？



# 制約と現実を忘れない

- 実装・運用制約
  - モデル更新頻度: 毎週? 毎レース? 更新コストは?
  - コード量・メンテナンス性、バージョン管理の手間
- 法的・倫理的制約
  - 公営ギャンブルの規定 / 未成年利用不可
  - オッズ操作や情報格差を生む行為は NG
- 結論
  - 良い問い = 実行可能な問い
  - ゴール・指標だけでなく“現実の壁”を先に洗い出しておく



# 良い問いのチェックリスト

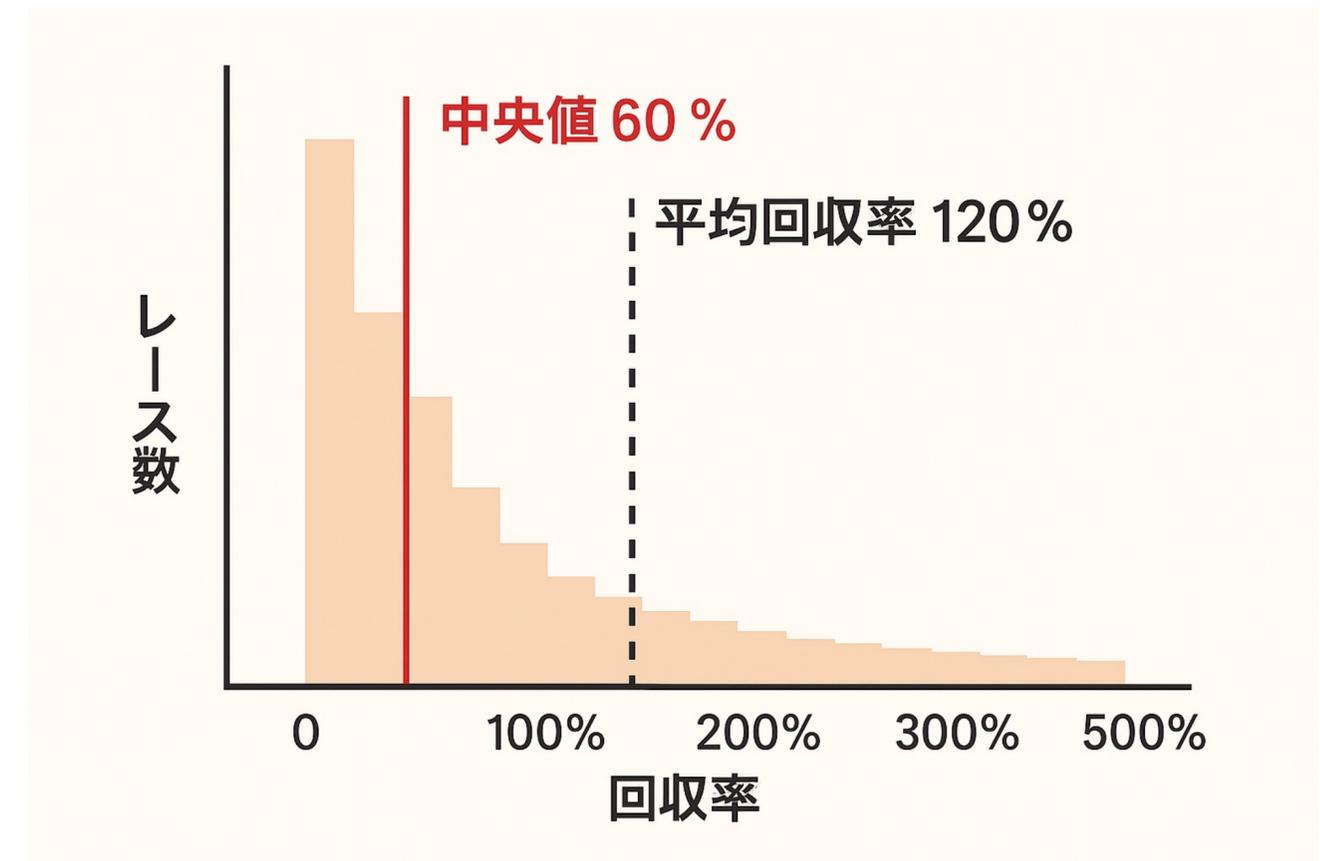
- ☑️ ゴールが具体・数値化されているか？
  - 「単勝的中率 50 %」「回収率 110 %」のように明確なターゲットを設定
- ☑️ 評価指標が問いに一致しているか？
  - 的中率／ROI／RMSE など、ターゲットごとに最適な指標を選定
- ☑️ 制約(時間・データ・資金)が含まれているか？
  - 発走前データのみ、スマホ推論 1 分以内、1 日 1,000 円以内…など
- 🎯 **Ask** → Check → Decide
  - ターゲット(何を当てる?) + 評価指標 + 制約 をセットで言語化
  - “Why before How、——目的が曖昧だと後工程がすべて崩れる”

# Check — データを疑う

回収率120%？

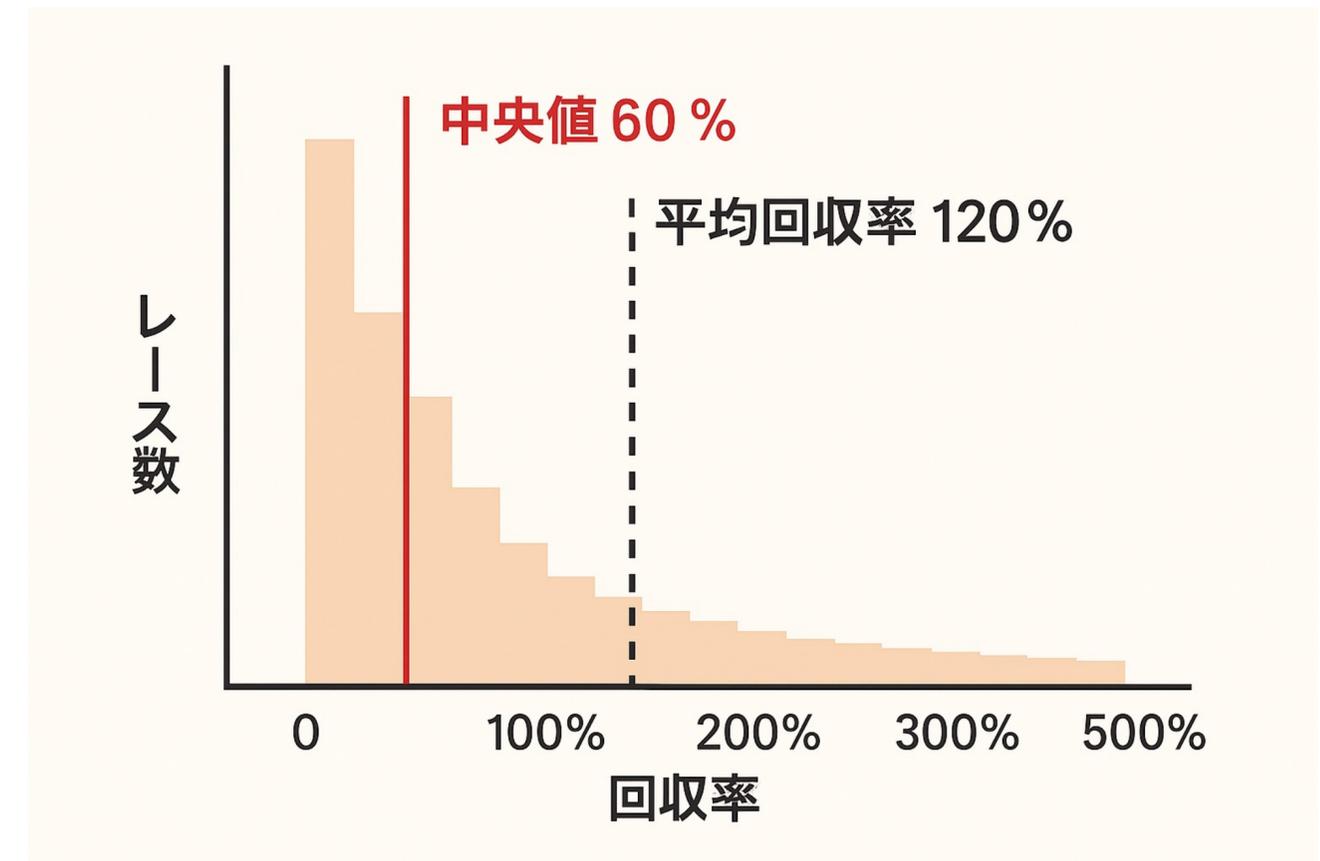
# 平均の罠：勝っているはずが赤字？

- “悪い報告書”の数字
  - 平均回収率(ROI) = 120 % と強調
  - しかし 中央値は 60 % → 半分以上のレースで赤字
- 右に長い分布
  - 万馬券がごくまれに当たり、平均を押し上げている
  - 大半は小額損失で“じわじわマイナス”



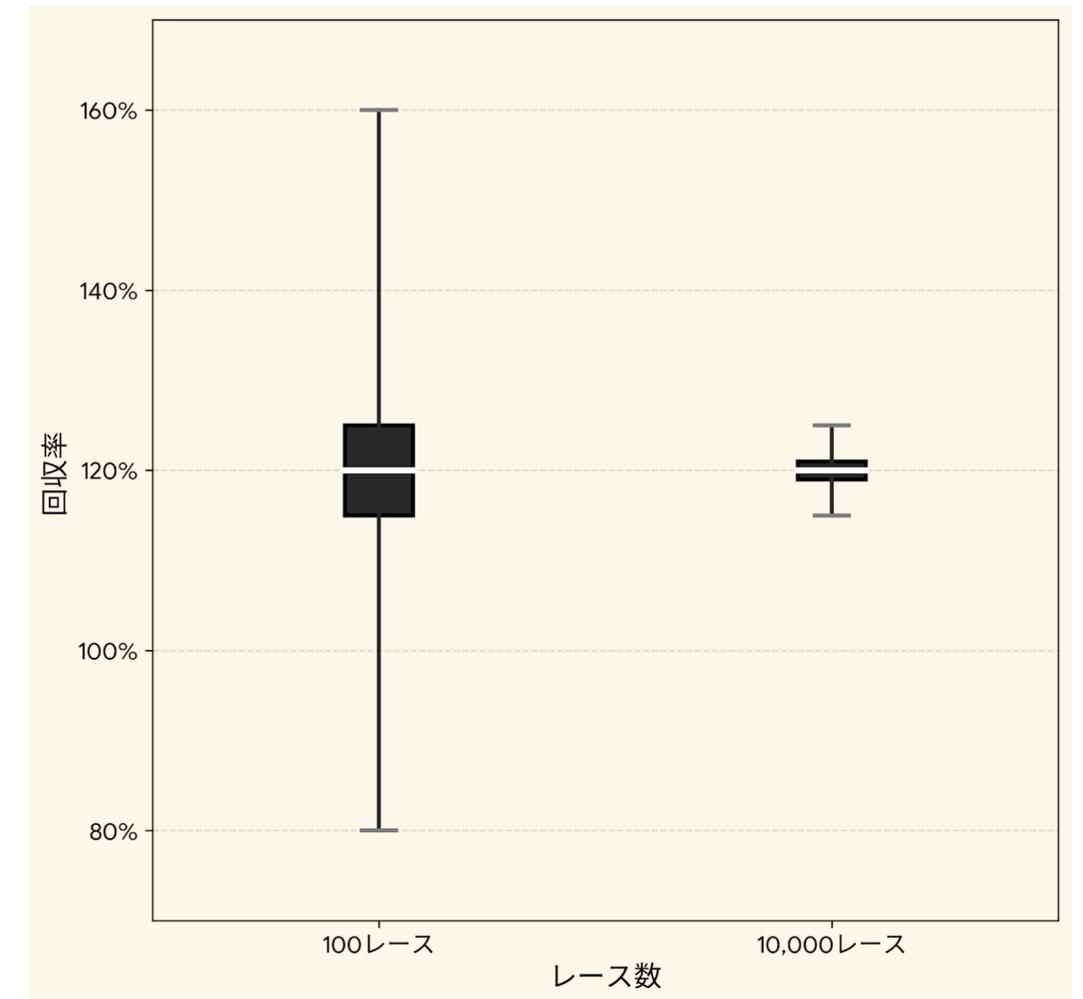
# 平均の罠：勝っているはずが赤字？

- 平均だけを見る危険
  - 「勝っている」と誤解し資金を増やす → さらなる損失に
- 確認すべき指標
  - 中央値・分散・箱ひげ図 で散らばりを可視化
  - 頻度表 で黒字／赤字レース数をカウント



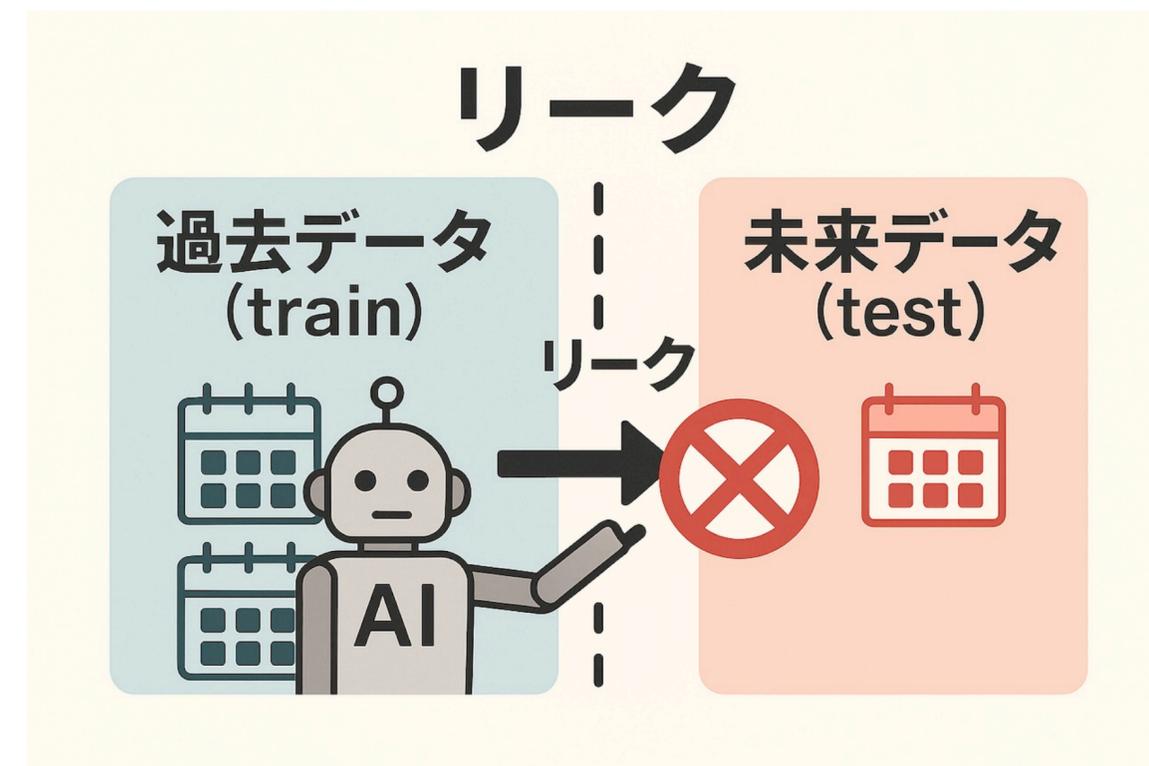
# サンプルサイズの罨：たった100レースで語る？

- 小規模データは“運”が支配
  - 100 レースでは偶然の当たり外れで ROI が  $\pm 40\%$  もブレる
  - → “再現性ゼロ” の数字をつかんでしまう危険
- 信頼区間で不確実性を可視化
  - 100 レース：95%信頼区間 が  $\pm 40\%$
  - 10,000 レース：95%信頼区間 が  $\pm 4\%$  に収束
  - n が 100 倍なら誤差幅は  $\frac{1}{\sqrt{100}} = \frac{1}{10}$  になる



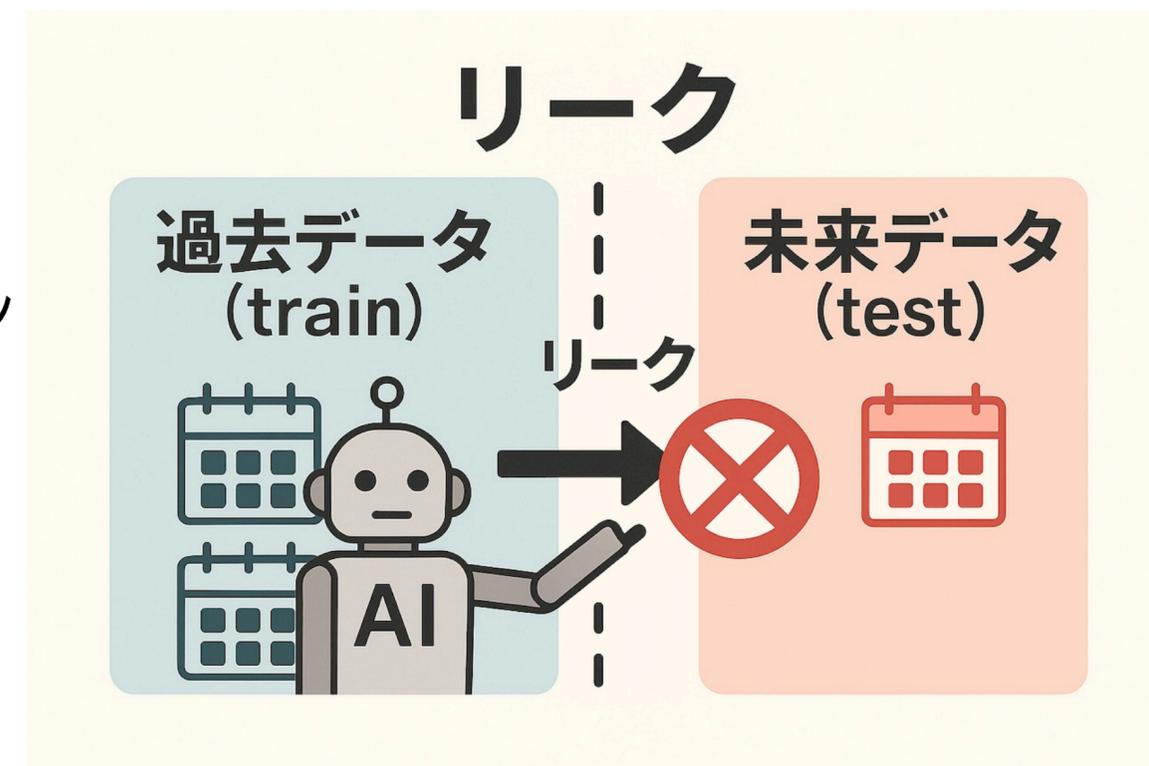
# リーク & 未来情報の混入

- リークとは？
  - 学習時に“未来でしか分からないデータ”がモデルに入るカンニング行為
  - オフラインでは高精度、本番では即崩壊
- 競馬でありがちなリーク例
  - 最終オッズ を特徴量に使用 → 発走前には未確定
  - レース結果(順位・タイム差)を推定値として逆算
  - 当日 馬場状態・降雨量 を事前に固定値で入力



# リーク & 未来情報の混入

- リークの影響
  - 的中率・RMSE が「ほぼ完璧」に見える
  - しかしライブ運用ではパフォーマンスが激減 → 損失増
- 防止策
  - 時系列分割(train: 過去 → test: 未来)でクロスバリデーション
  - 発走前に取得できない列は物理的に除外
  - ライブシミュレーション(当日データを順次入力)で再検証
  - 重要特徴量を確認し「未来情報」が混ざっていないか点検



# Quick Check List: データにツッコミを！

- ✔ 平均だけでなく中央値・分散も確認 → “右に長い尻尾”は平均を押し上げる → 箱ひげ図で散らばりを把握
- ✔ サンプル数と信頼区間をセットで示す →  $n$  が少なければ  $\pm$ 誤差幅を明記 / 時系列 CV で再検証
- ✔ 未来で見えない列が混入していないか？ → 最終オッズ・レース結果など “カンニング列” を物理的に除外
- ✔ グラフで可視化し“変だ”をあぶり出す → ヒストグラム・散布図で外れ値やリーク臭を即発見
- ✔ 数字を鵜呑みにせず必ず“なぜ？”を一回問う → 高精度・高 ROI はまず疑う → ソースと手順を再チェック

🎯 Ask → Check → Decide

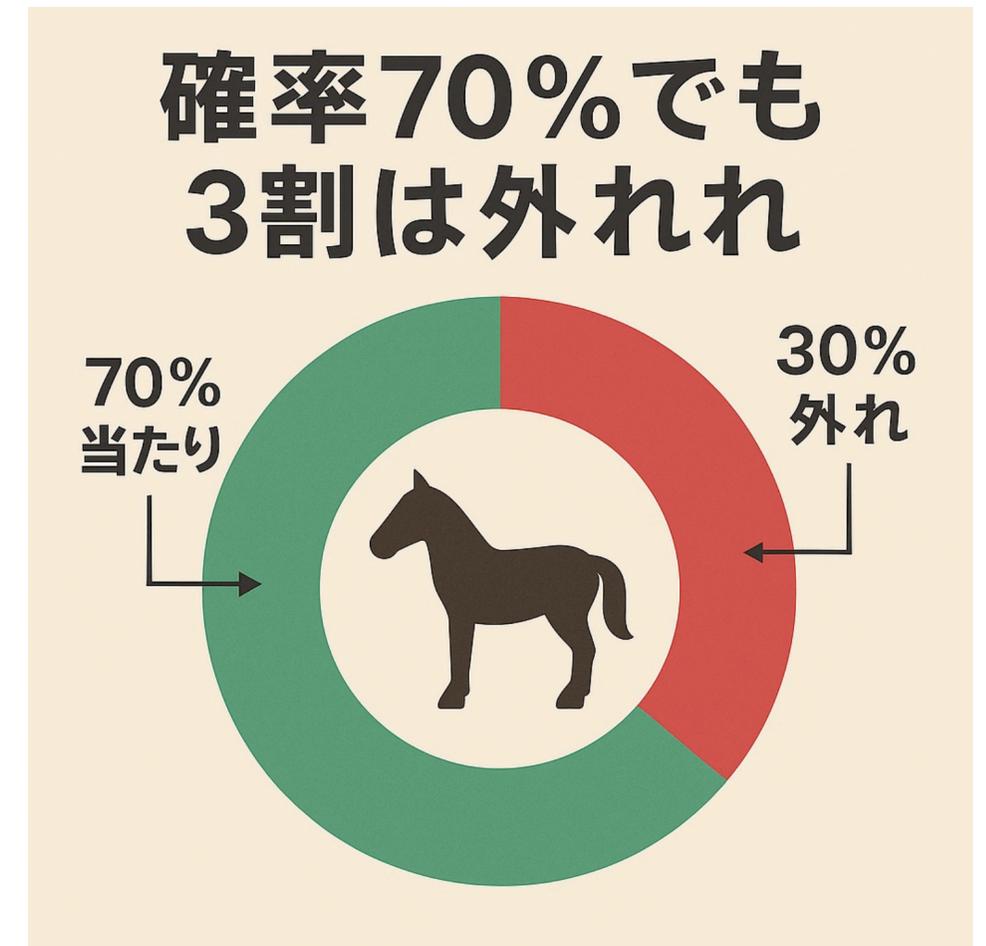
- “Check” を通さずに出した数字は 信じない のが鉄則。

# Decide — 予測を”使える知恵”に変える

予測を活かす、決断を下す

# 理解する: 確率を正しく読む

- 確率 = 保証ではない
  - 的中率70 % は「長期平均で当たる割合」
  - 各レースはコイン投げ同様、外れる可能性 30 % が常に残る
- 期待値と実現値のギャップ
  - 10 レース中 7 勝 3 敗は “あくまで目安”
  - 長期的に収束する勝率7割に収束する
- 資金管理の必要性
  - 資金を賭ける割合を固定 → 破産確率を抑える
  - ケリー基準などで「外れた30 %」に耐える設計



# 理解する：現実の不確実性

- データに無い“突発要因”
  - 急な雨・馬場悪化／ゲート再装填・出遅れ
  - 馬の体調変化・騎手の戦術変更 など
- 残差はゼロにならない
  - モデルは“過去に説明できた範囲”まで
  - 常に 不可避なノイズ が残る＝外れの源
- モニタリング&再学習
  - 精度低下を検出→モデル再訓練／特徴量追加
  - 天候・馬体重など“当日変数”をリアルタイム注入

**モデル外要因：  
現実揺らぐ**



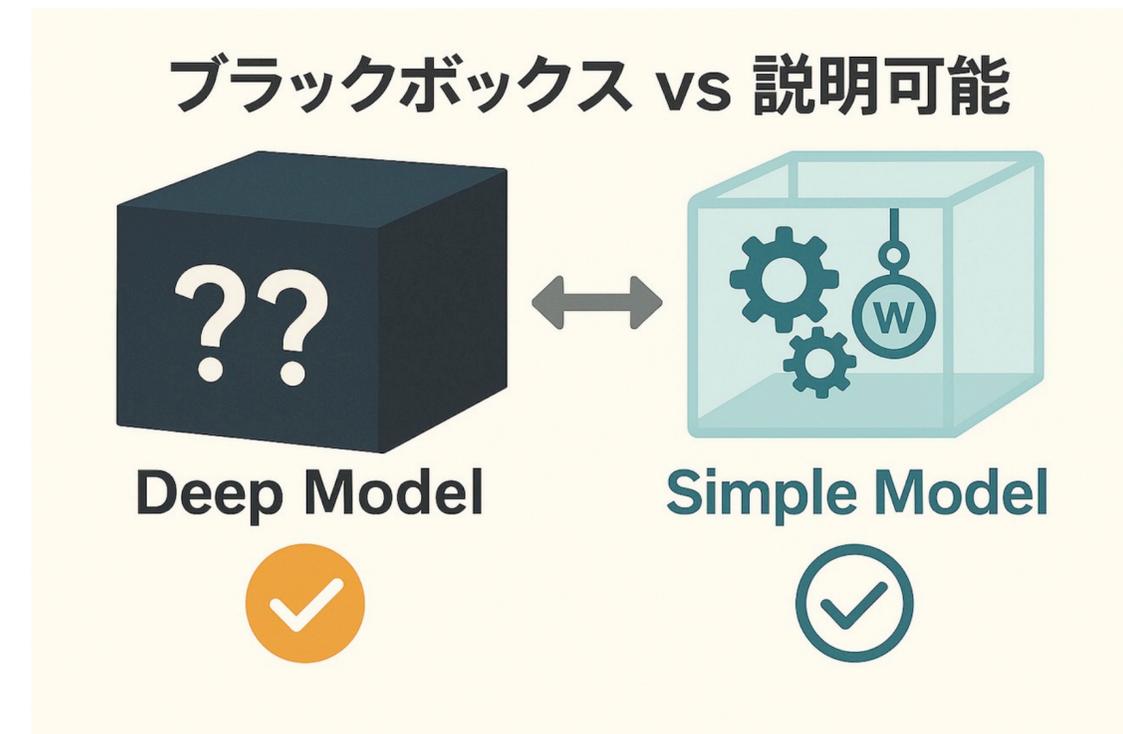
# 選ぶ: モデルを賢く選ぶ

- 深層モデル・ディープラーニング(ブラックボックス)
  - 高精度・多特徴量を自動抽出
  - しかし“なぜ外れた?”を説明しづらい
  - 再学習コストが高く、推論も重い
- シンプルモデル(ロジスティック回帰など)
  - 重み係数で要因が一目瞭然
  - 精度は深層モデルより落ちやすい
  - 学習・推論が速く、実装も容易

観点	深層モデル	シンプルモデル
精度	◎ 高い	○ 中程度
説明性	△ 低い	◎ 高い
計算コスト	△ 高い	◎ 低い
デバッグ容易さ	△ 難しい	◎ 容易

# 選ぶ: モデルを賢く選ぶ

- 規制・信頼性の観点
  - ギャンブル依存対策やユーザー説明責任 ⇒ 説明性が求められる場面あり
  - 逆に高速自動売買などは精度最優先でブラックボックス容認も
- メッセージ
  - 「当てる力」だけでなく「納得させる力」も選択基準に
  - 目的・制約に応じてモデルの バランスを設計 することが重要



# 決める: 自分の基準で決断

- 同じ予測でも賭け方で結果が激変
  - 的中確率30%・回収率120%でも“賭け額”次第で破産も黒字も
- ベット戦略の比較
  - ケリー基準 — 資金成長を最大化（高リスク・高リターン）
  - 固定額 — 収益は小さいが変動も小さい
  - Do Nothing — 賭けない選択肢も戦略
- 自分のリスク許容を数値化
  - 1レースの最大損失／月次ドローダウン／目標利益…を先に決める



# 意思決定のチェックリスト

- ☑ その予測は確率として妥当か？
- ☑ モデル外の要因を見落としていないか？
- ☑ なぜそう判断したのか説明できるか？
- ☑ 自分のリスク許容に合っているか？
- 🎯 Ask → Check → Decide
  - 最終判断はモデルではなくあなたの価値観と制約で下す

# 本日の講座まとめ

- 競馬AIを通じて体験したこと
  - AIは「問い → データ → 判断」をつなぐ “確率の道具”
- ① **Ask**: 良い問いを作る → 自分の興味領域で “問い” を数値化してみる
  - ターゲット + 評価指標 + 制約をセットで数値化
- ② **Check**: データを疑う → 最低 1 つグラフを描き、データにツッコミを入れる
  - 平均の罨・サンプル不足・リークをチェックリストで回避
  - 可視化で “変だ” を炙り出す
- ③ **Decide**: AIの限界を踏まえ意思決定 → 結果をどう使うか責任を持って判断する
  - 70 %でも3割外れる — モデル外要因と確率の揺らぎを許容