

データサイエンス概論I & II データサイエンス総論I & II

予測と回帰分析

九州大学 数理・データサイエンス教育研究センター

データを用いた「予測」

だれでも予測をしながら生きている。
でも正直、正確な予測は、現代でも難しい

身近な予測①

未来の予測

- 試験の勉強

- 過去の傾向からみて、明日はきっとこの問題が出るだろう



- スポーツ

- 次はストレートを投げてくるに違いない



- 買い物

- この値段・素材のものを買えば、5年は大丈夫だろう



- 天気予報

- 過去の天気データを用いて、明日以降の天気を予測



他にも...

- 株価の予測
- 競馬等のギャンブル
- 就職活動 などなど



身近な予測②

未来ではなくても、「**だろう**」がつけば全部「予測」

● 画像認識

- (無意識に)「この動物は犬だろう」
- (無意識に)「あ、機嫌が悪そうだ」
- この本（表紙とタイトル）は、きっと面白いだろう



「見た目」（画像）がすべてではない。
それでも我々は（ある程度正確に）
「これは何だ！」と予測して生きている！



● 推量・診断

- これぐらい勉強すれば、これぐらいの点数は取れるだろう
- この体温ならば、インフルエンザだろう



● 因果推論 (= こういう結果になったのはこういう原因があったからだ)

- 警察の推理, 故障原因の推定, 考古学



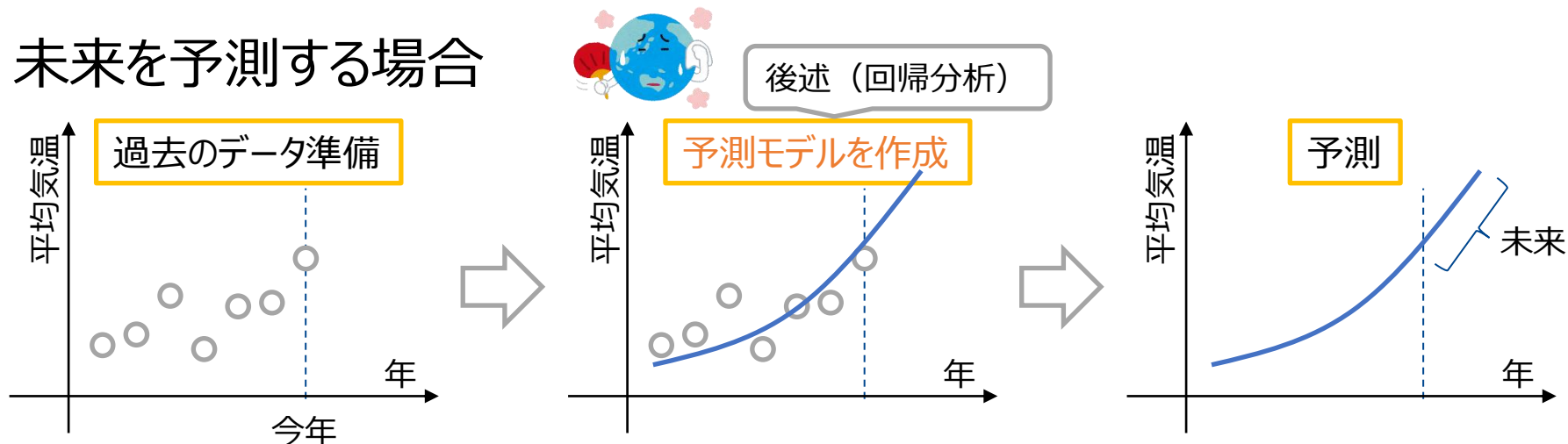
● 推薦

- このユーザーなら、この商品なら買ってくれそう！

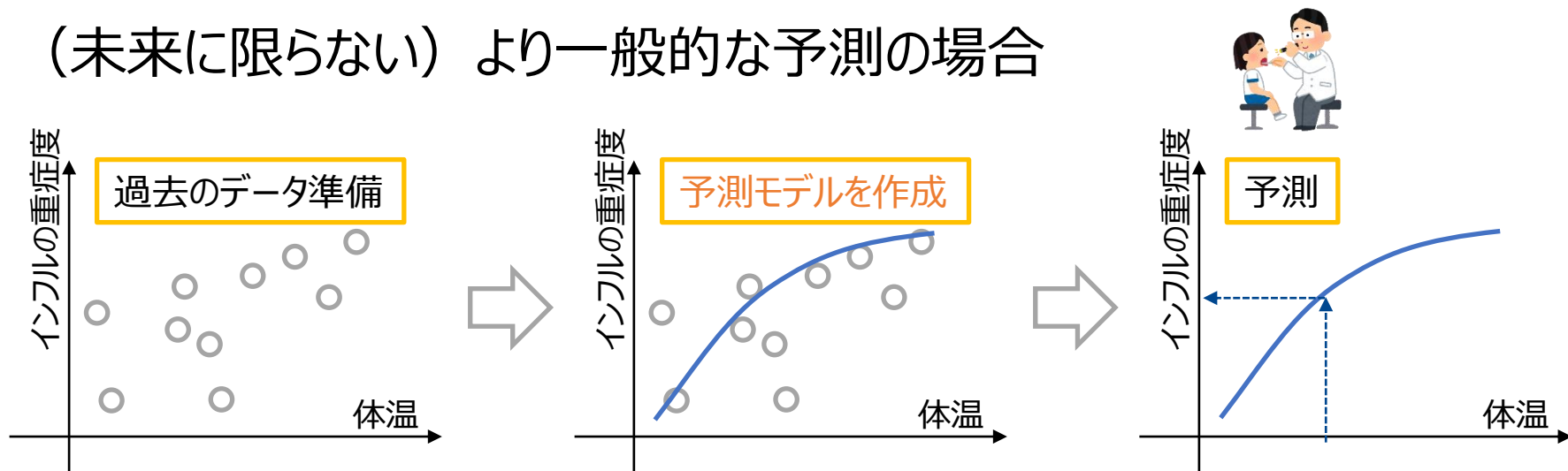


データを用いた予測の方法： 難しそうに感じるかもしれませんが、みんな無意識にやっていることです

● 未来を予測する場合

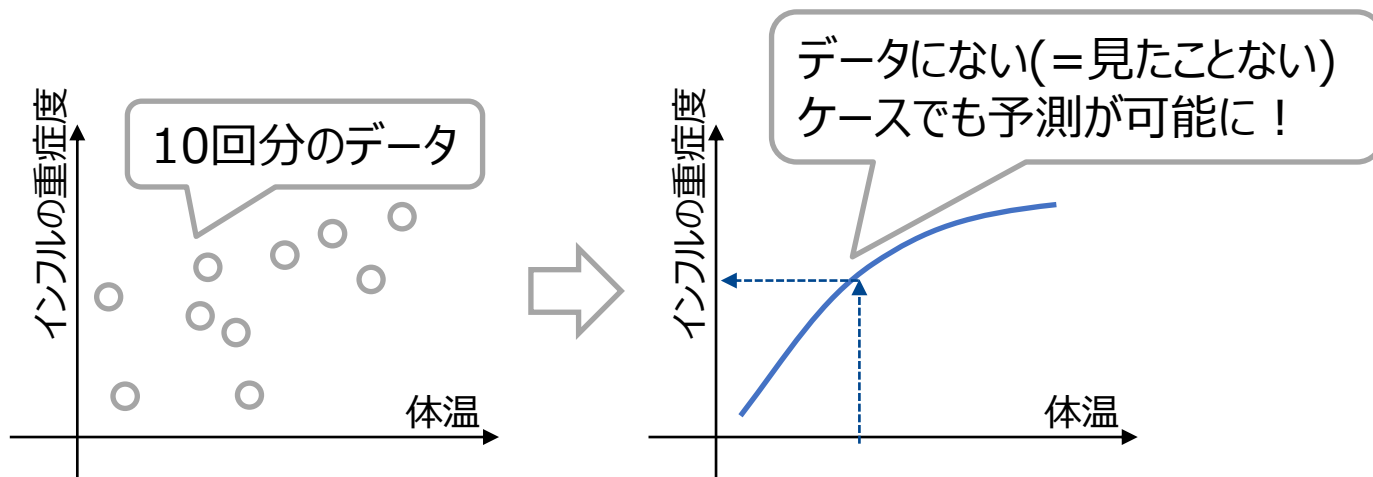



● (未来に限らない) より一般的な予測の場合



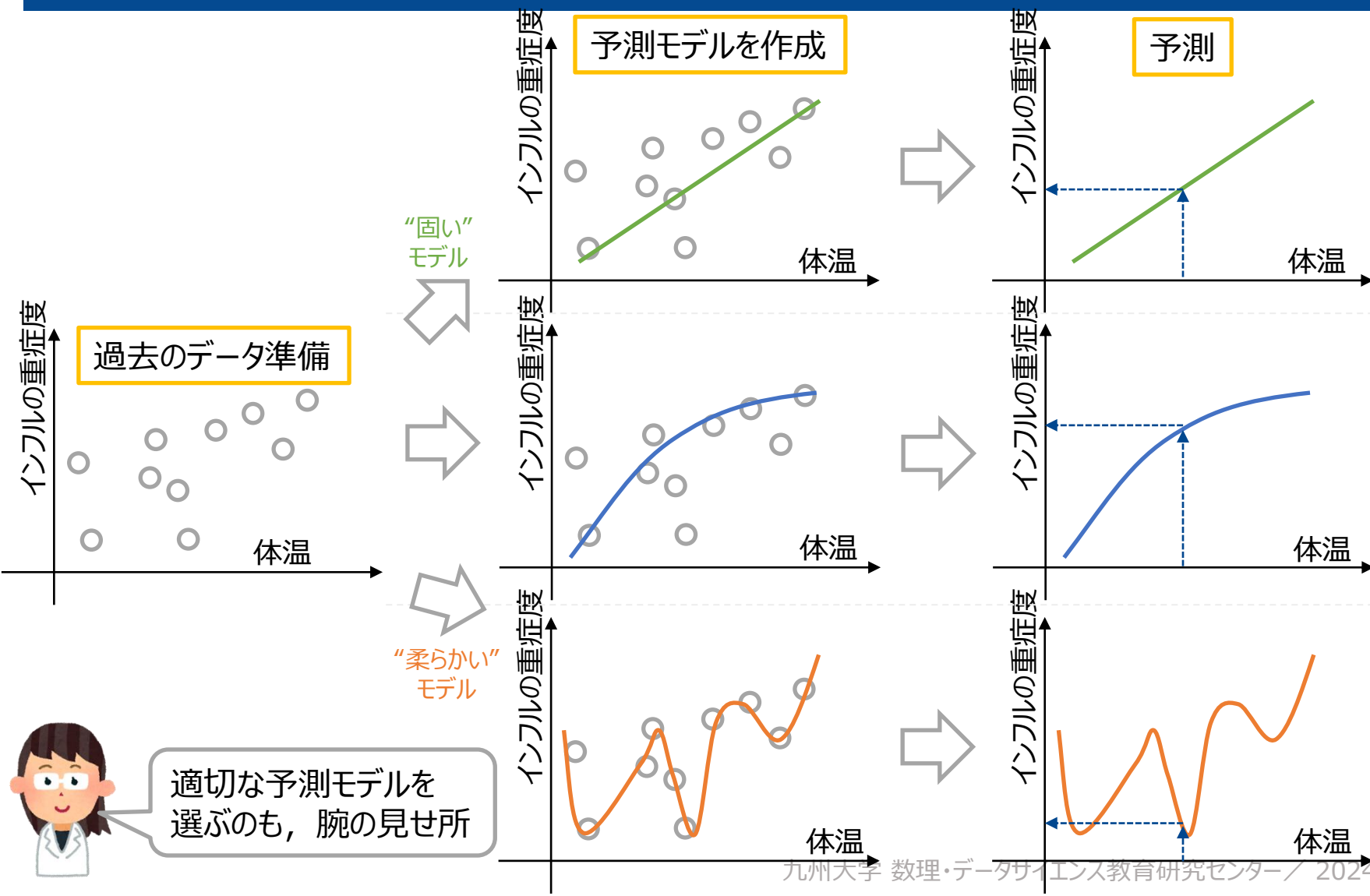
予測モデルのうれしいところ

- 過去になかった状況に対しても予測可能



- 我々人間も、過去の数回の経験(=データ)に基づいて、未知の状況でも何らかの予測を行いながら生きてる！
 - 「以前こうなったから、今回はたぶんこうなる」という知識が皆さんの「予測モデル」
 - 例：医師も、他の患者の診察結果に基づいて、初診の患者を診察している
 - 例：見たことのないタイプの犬でも、過去に見た犬に基づき、犬とわかる

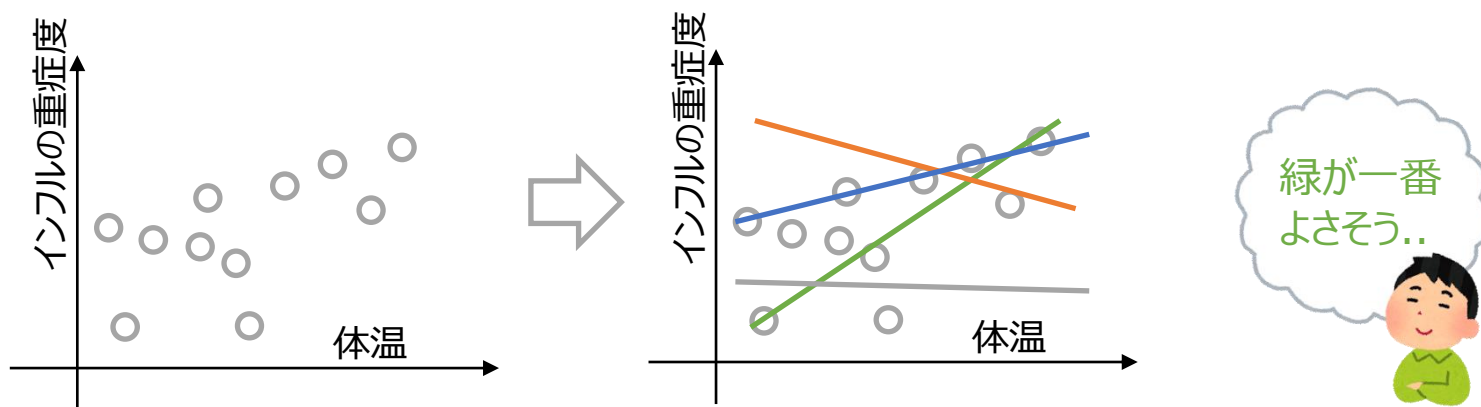
予測モデルは様々考えられる。そして、 モデルによって予測結果は異なる(精度が違う)



さらに…

同じ予測モデルでも、「あてはめ方」は色々

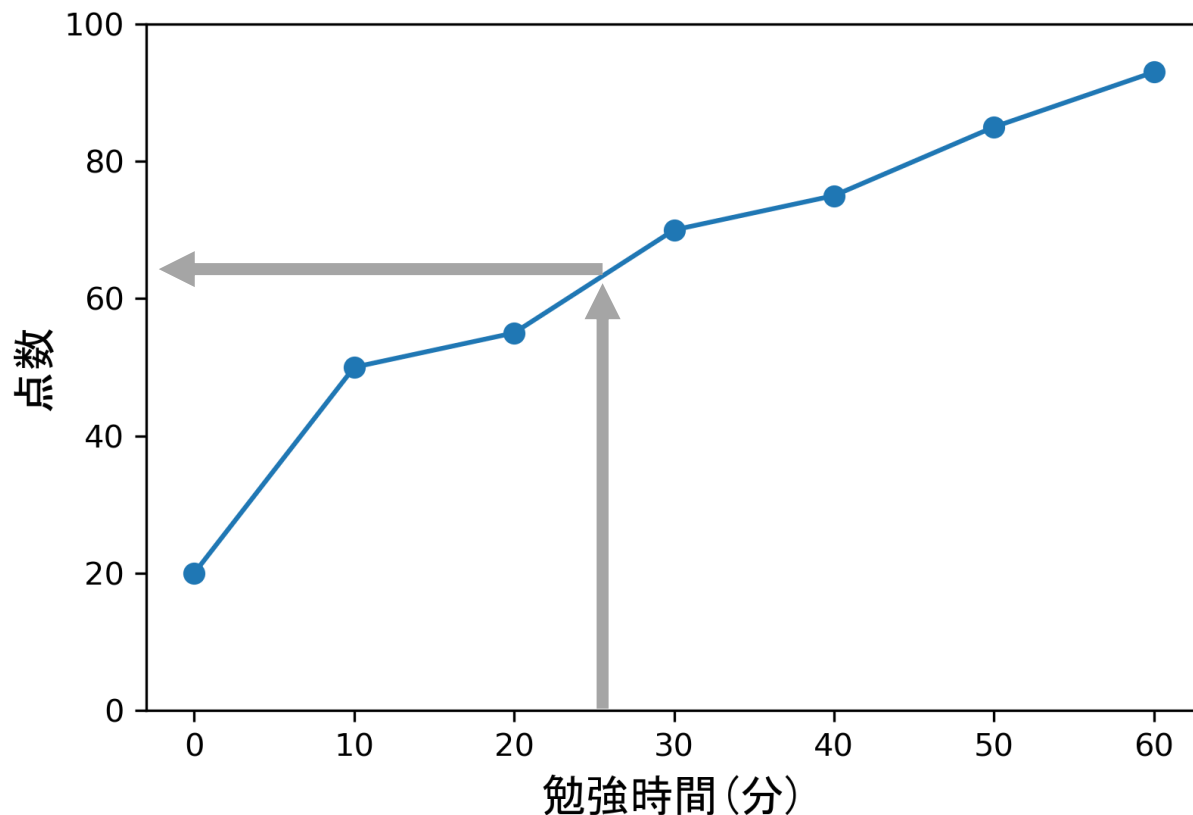
- 例えば同じ「まっすぐな予測モデル(線形予測モデル)」でもデータへの「あてはめ方」は色々考えられる



- なるべく多くのデータを正確に予測できるように、適切にあてはめる必要がある
 - 後述する「最小二乗法」が一般的

ちなみに、小学生でも予測をやっている！ 折れ線グラフ、実は最も単純な「予測」

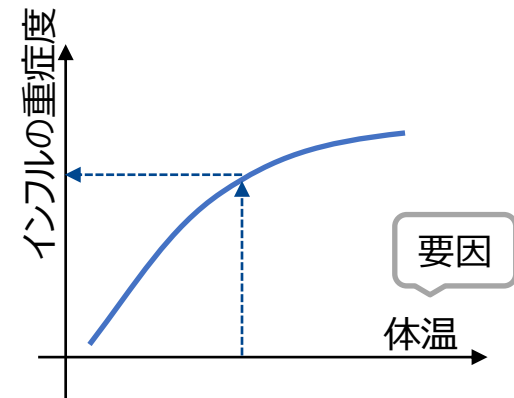
- 予測モデル = 折れ線



- 未測定の勉強時間(ex. 25分)でも点数を予測可能(63点ぐらい)

しかし予測は難しい！

- 過去のデータを十分に集められない場合がある
 - 「あなたの10年後の給料」を予測するためには、「あなたと似たような人」をたくさん集める必要がある
- 予測結果を決める要因が不明な場合がある
 - インフルの重症度は体温だけでよいのか？
 - 上記の「10年後の給料」予測に必要な要因は？
 - 天気予報のように要因がほとんど無限に存在する場合もあり
- 現時点と予測時点では状況が違う場合がある
 - 2年後に突然不況が起こったら、「10年後の給料」予測結果は外れる
 - = いつまでも同じ予測モデルが使えるとは限らない
- どの予測モデルを使えばよいかは、自明ではない
- … など様々な難しさがある！



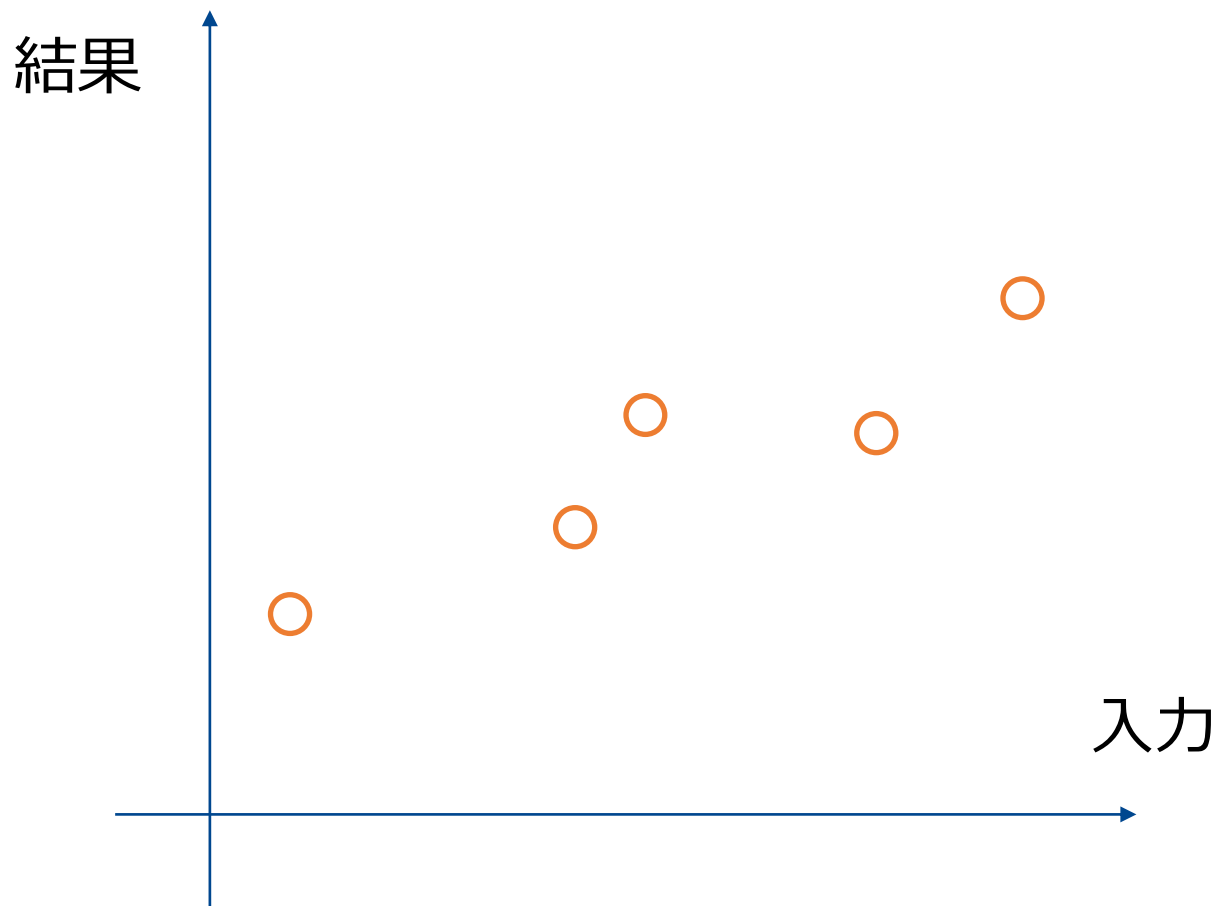
専門家でも、
予測は難しい



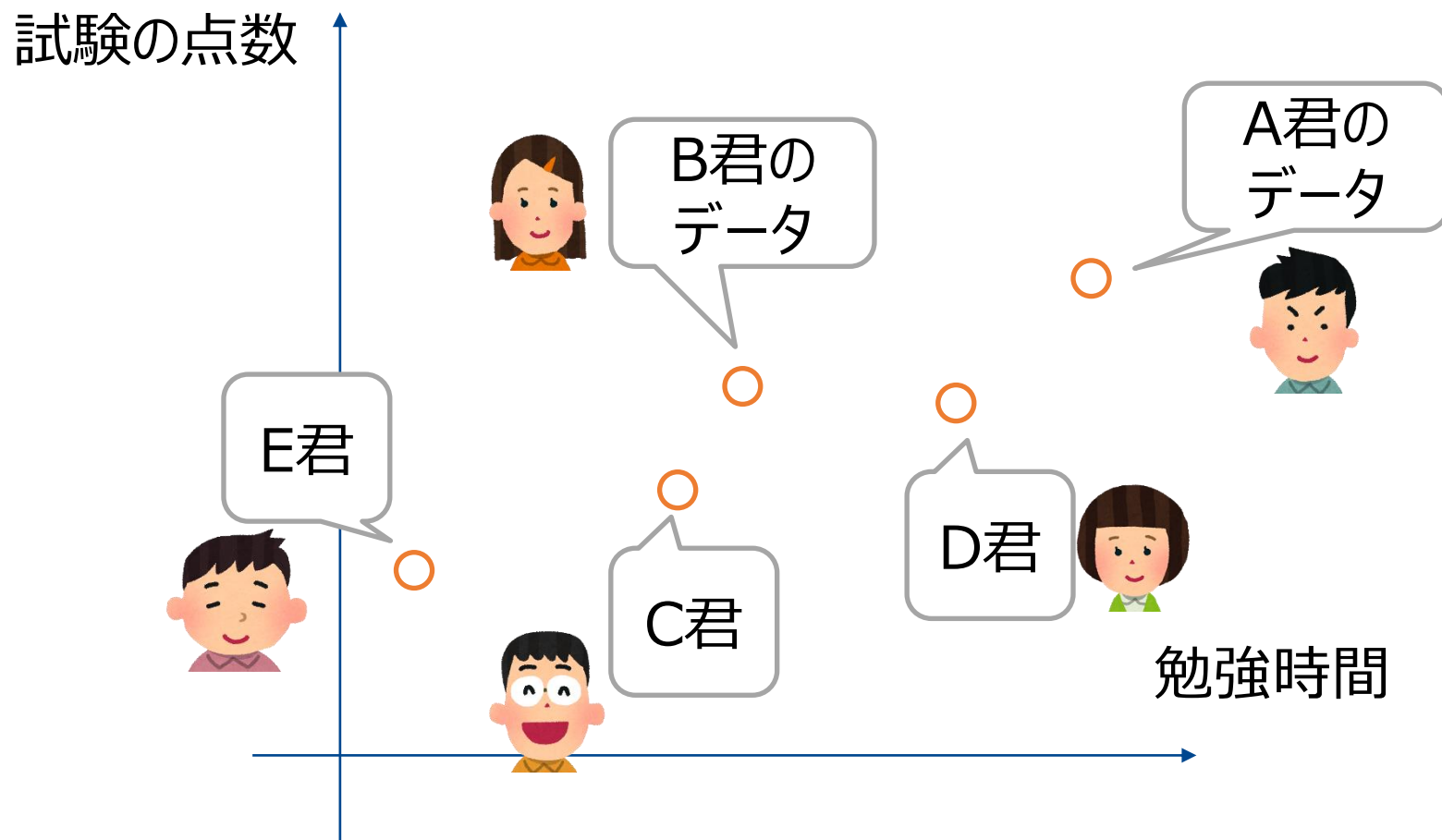
回帰による予測

すでにお伝えしたことをもう一度

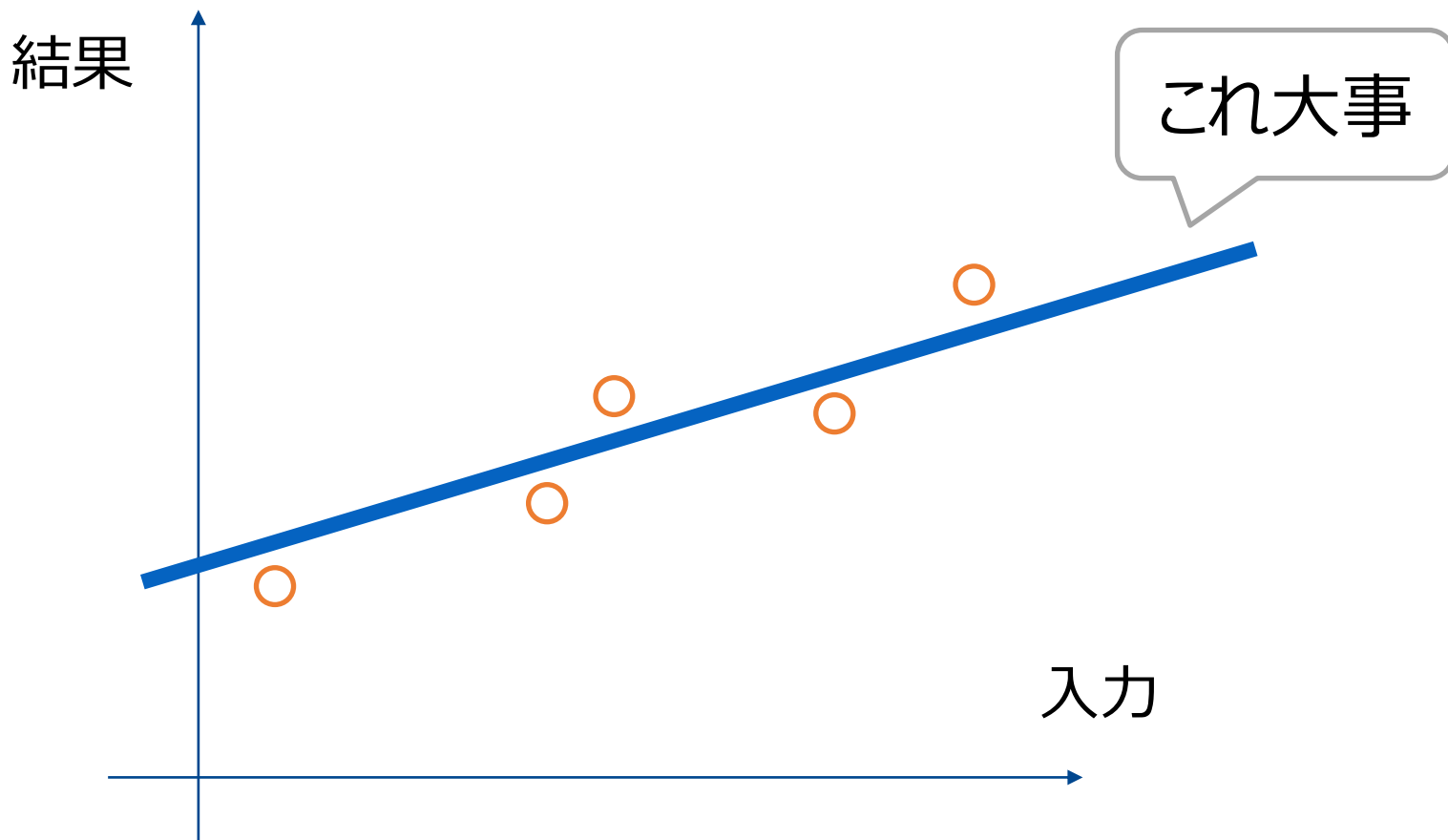
回帰による予測（ステップ1/3）： データ収集



回帰による予測（ステップ1/3）： 「データ収集」for 勉強時間から点数を予測

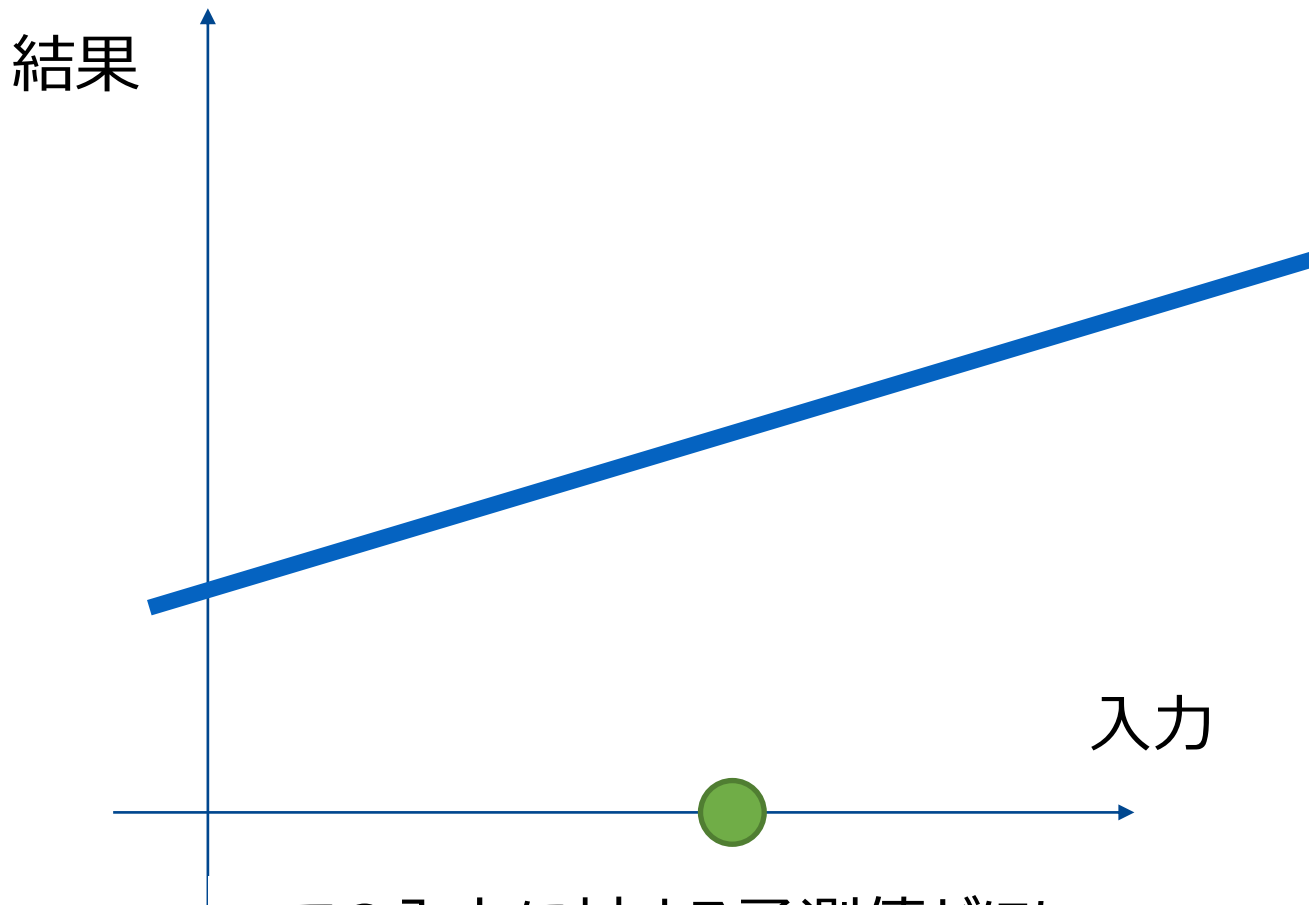


回帰による予測（ステップ2/3）： モデルあてはめ



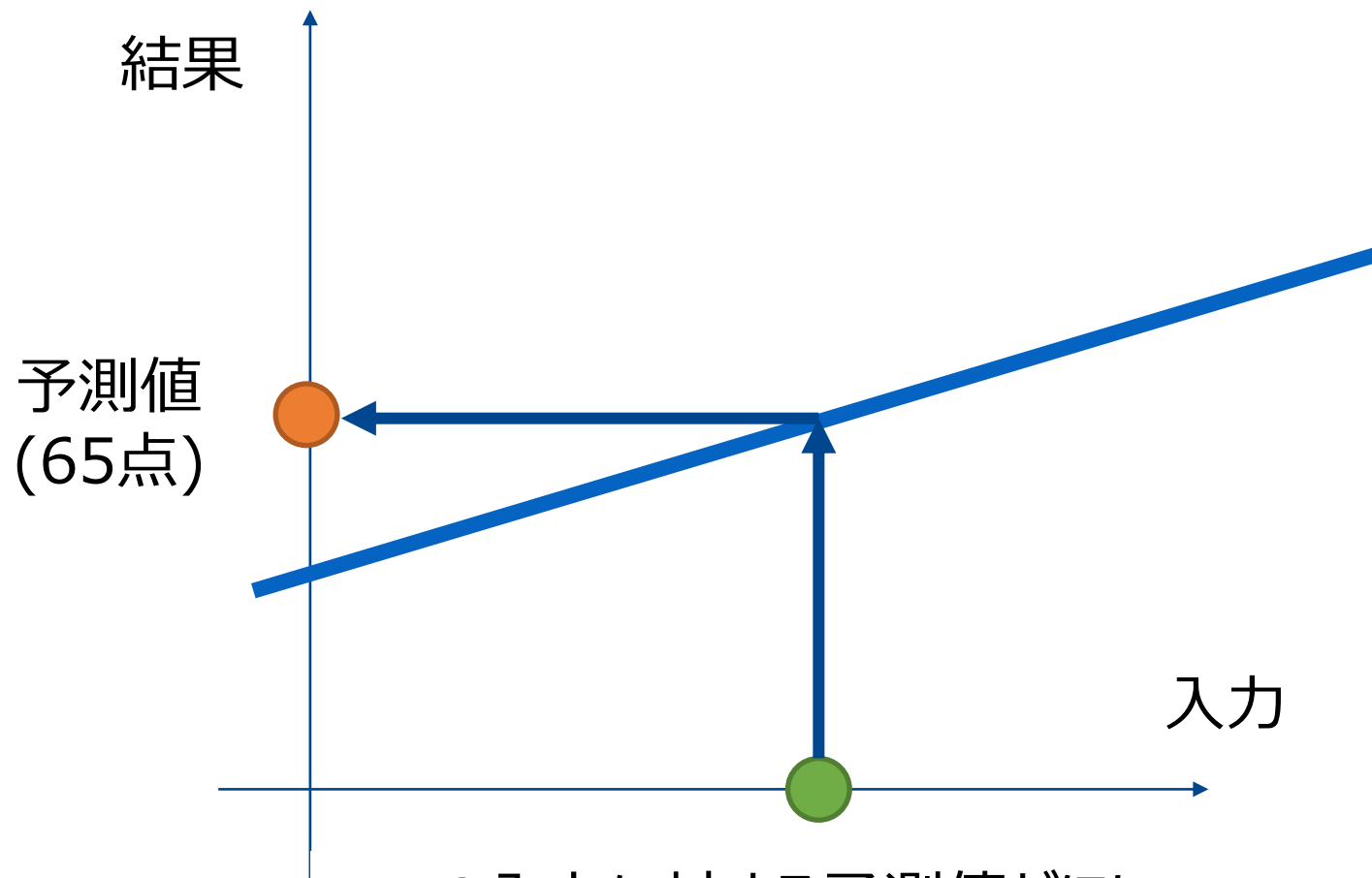
モデル = 「条件 or 入力」と「結果」間に成り立つと予想される関係。
上記は「線形モデル」

回帰による予測（ステップ3/3）： 予測



この入力に対する予測値がほしい
(ex. 勉強3時間だと何点取れそう?)

回帰による予測（ステップ3/3）： 予測



この入力に対する予測値がほしい
(ex. 勉強3時間だと何点取れそう?)

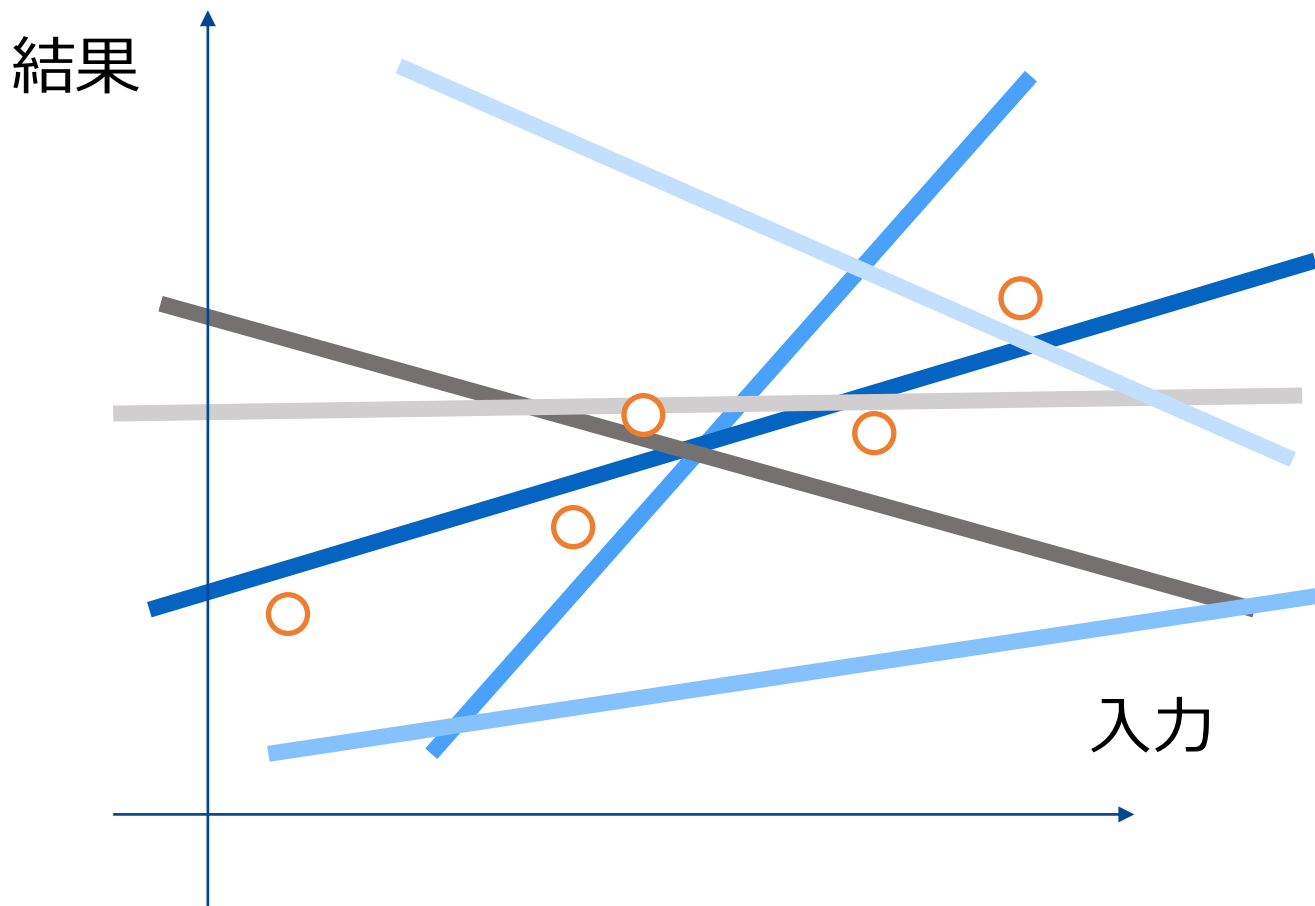
回帰分析とは

- いくつかの（入力，結果）の事例に基づいて，
- それらの関係をモデル化することで，
- 新たな入力に対して，その結果を予測する方法．
- 専門的な用語としては
 - 入力→説明変数
 - 結果→外的変数
 - とか言ったりしますが…

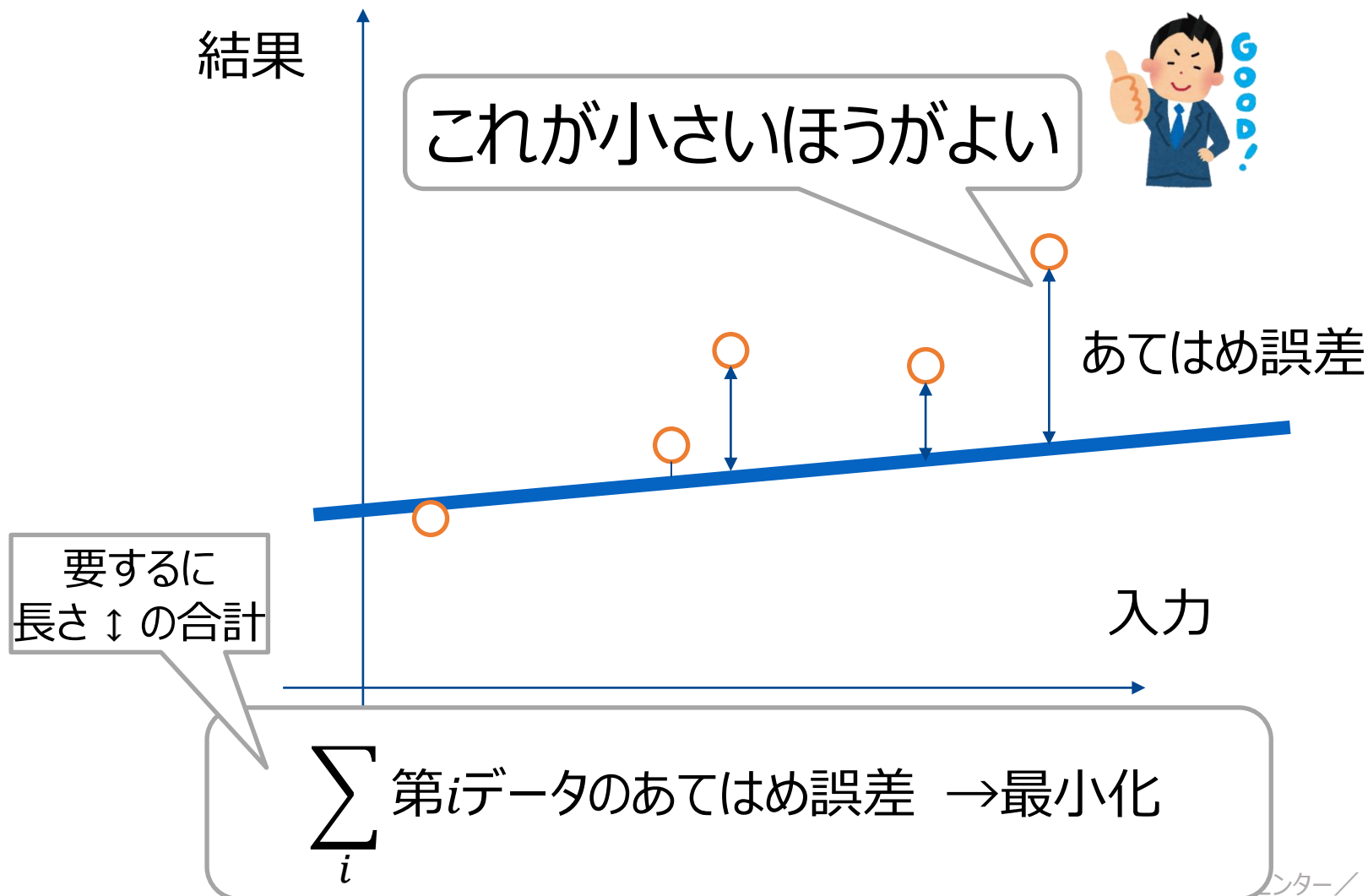
「モデルあてはめ」の方法

最小二乗法で解く！

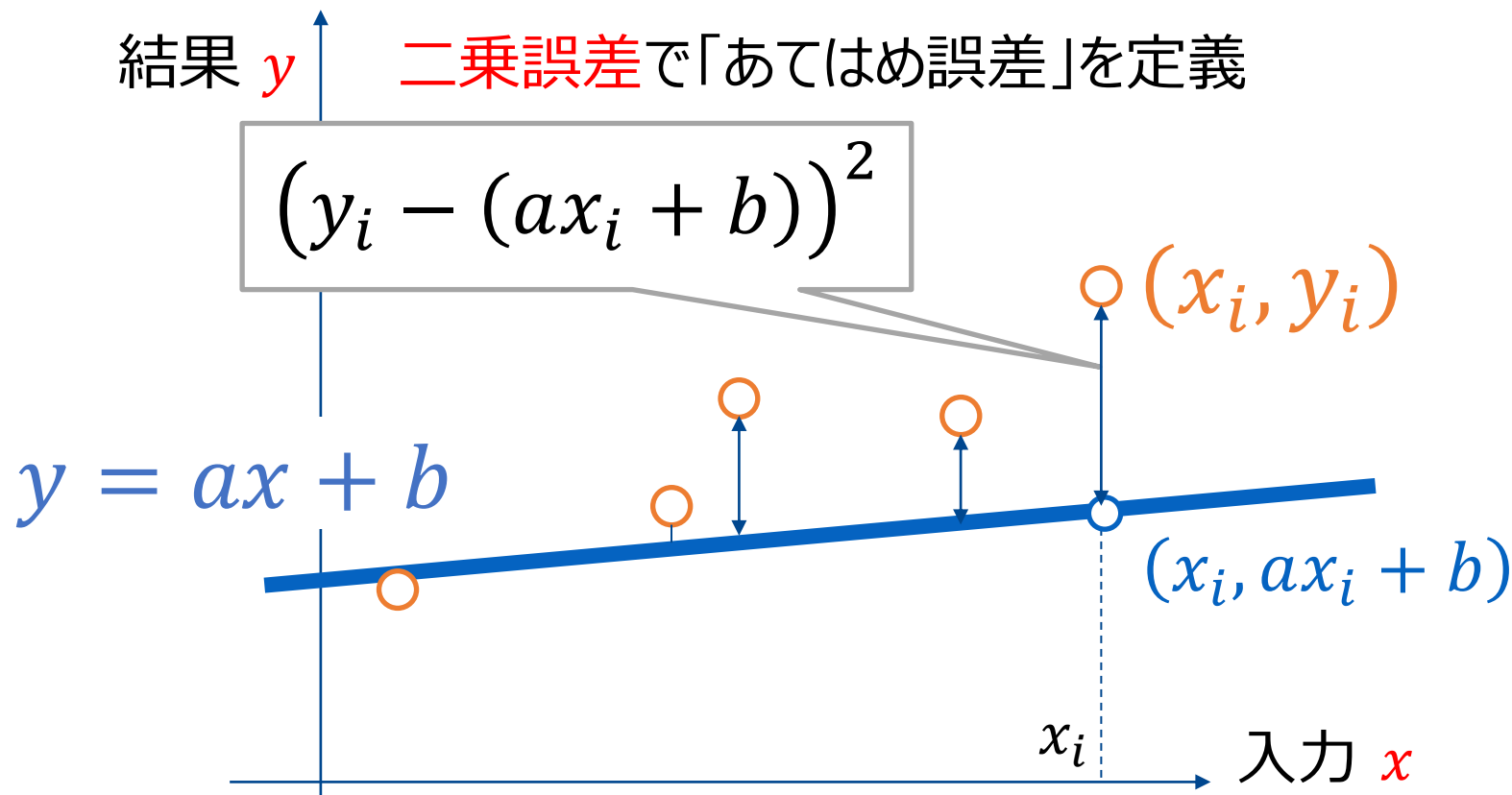
どうい「あてはめ結果」が望ましい？ (1/3)



どういう「あてはめ結果」が望ましい？ (2/3)

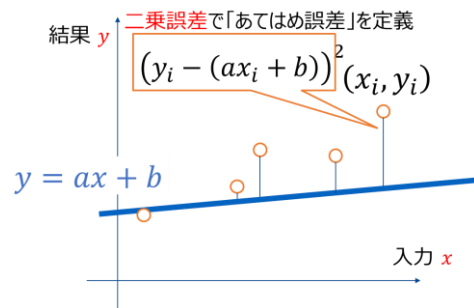


どうい「あてはめ結果」が望ましい？ (3/3)



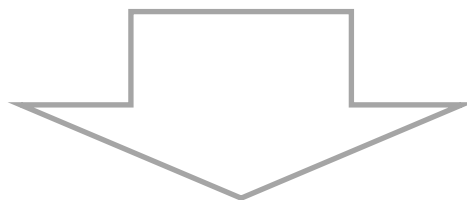
a と b をいじって $\sum_i (y_i - (ax_i + b))^2$ を最小化

これを「最小二乗法」と呼ぶ



二乗誤差を最小にしたいから
最小二乗法

a と b をいじって $\sum_i (y_i - (ax_i + b))^2$ を最小化



どうやって??



ちょっと頑張ってみよう：最小二乗法，どう解く？

まず問題をよく見る(1/3)

主役の一人 a

最小化したいもの

$$\sum_i (y_i - (ax_i + b))^2$$

$$= \sum_i (y_i - (ax_i + b))(y_i - (ax_i + b))$$

$$= \sum_i x_i^2 a^2 - (y_i - b) 2x_i a + (y_i - b)^2$$

落ち着いて考えれば
 a に関する2次関数

$$= \left(\sum_i x_i^2 \right) a^2 - 2 \left(\sum_i (y_i - b) x_i \right) a + \sum_i (y_i - b)^2$$

ちょっと頑張ってみよう：最小二乗法，どう解く？ まず問題をよく見る(2/3)

もう一人の主演 **b**

最小化したいもの

$$\sum_i (y_i - (ax_i + b))^2$$

$$= \sum_i (y_i - ax_i - b)(y_i - ax_i - b)$$

$$= \sum_i b^2 - (y_i - ax_i)2b + (y_i - ax_i)^2$$

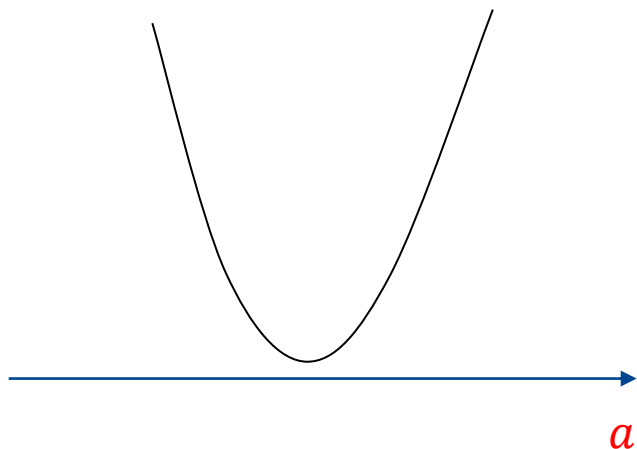
落ち着いて考えれば
 b に関する2次関数

$$= \left(\sum_i 1 \right) b^2 - 2 \left(\sum_i (y_i - ax_i) \right) b + \sum_i (y_i - ax_i)^2$$

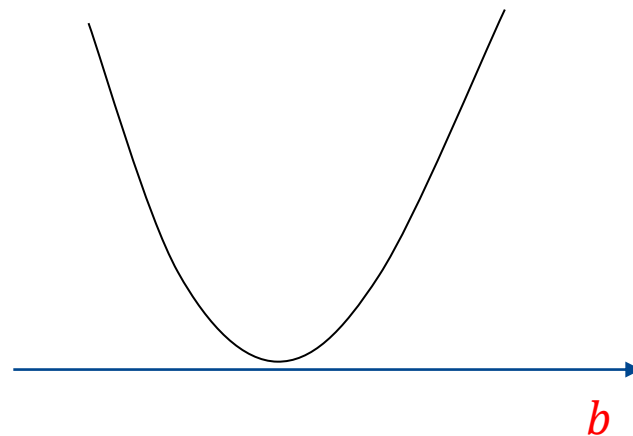
ちょっと頑張ってみよう：最小二乗法，どう解く？ まず問題をよく見る(3/3)

$$\sum_i (y_i - (ax_i + b))^2$$

a 方向から見ても2次関数



b 方向から見ても2次関数



ということは…



← 富士山はどこから見てもこんな感じ

ちょっと頑張ってみよう：最小二乗法，どう解く？

要するに誤差はこんな形になっています

誤差

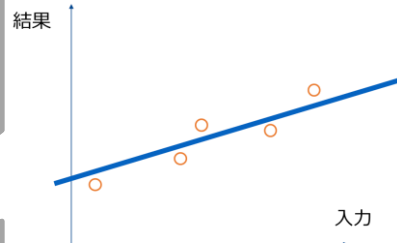
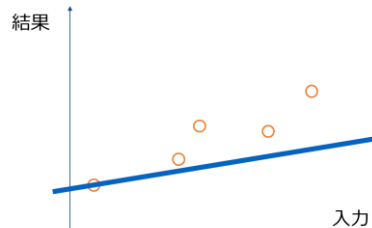
$$\sum_i (y_i - (ax_i + b))^2$$

誤差大

誤差最小

この (a, b) がベスト

この (a, b) はイマイチ



ベストは 1 か所 → ベストな「モデルあてはめ」は一つ！
(by 一か所だけ出っ張った二次関数の美しい特性)

ちょっと頑張ってみよう：最小二乗法，どう解く？

解き方の細かいところ

$$a \text{ と } b \text{ をいじって } \sum_i (y_i - (ax_i + b))^2 \text{ を最小化}$$



解き方(要は「極値なら微分してゼロ」という条件)

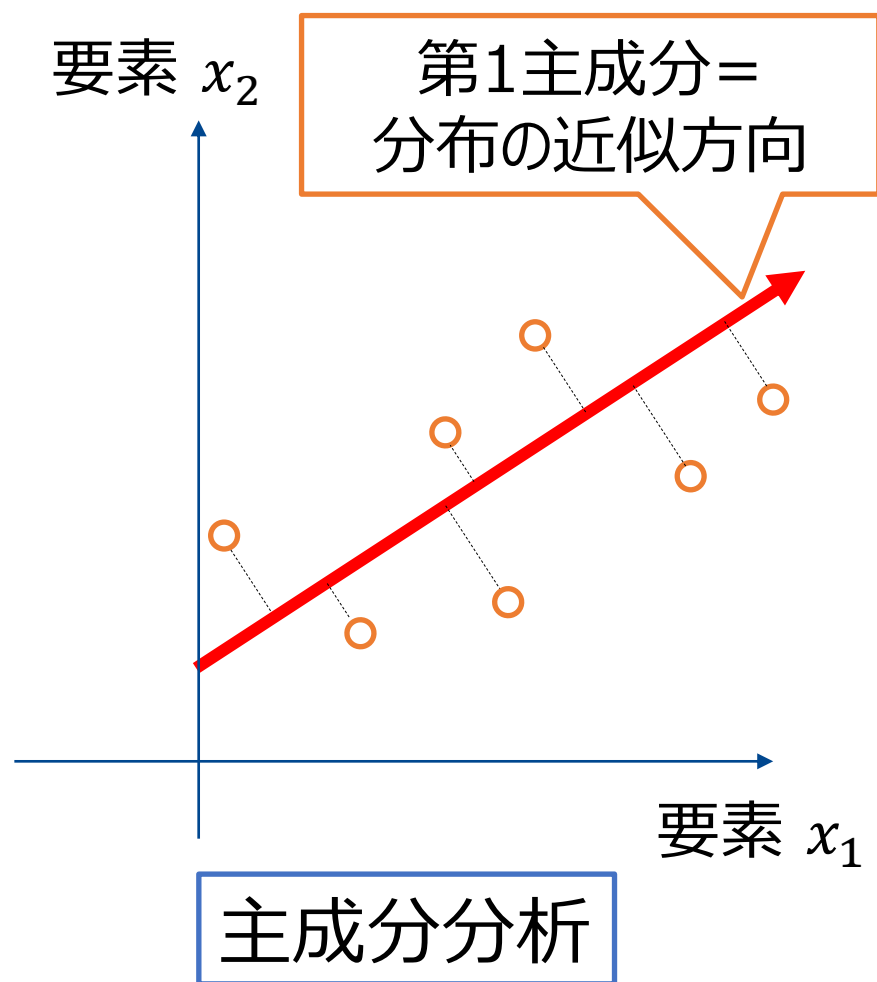
$$\begin{cases} \frac{\partial \sum_i (y_i - (ax_i + b))^2}{\partial a} = 0 \\ \frac{\partial \sum_i (y_i - (ax_i + b))^2}{\partial b} = 0 \end{cases}$$

高校時代
二次関数の
極値問題を
解かされたな...

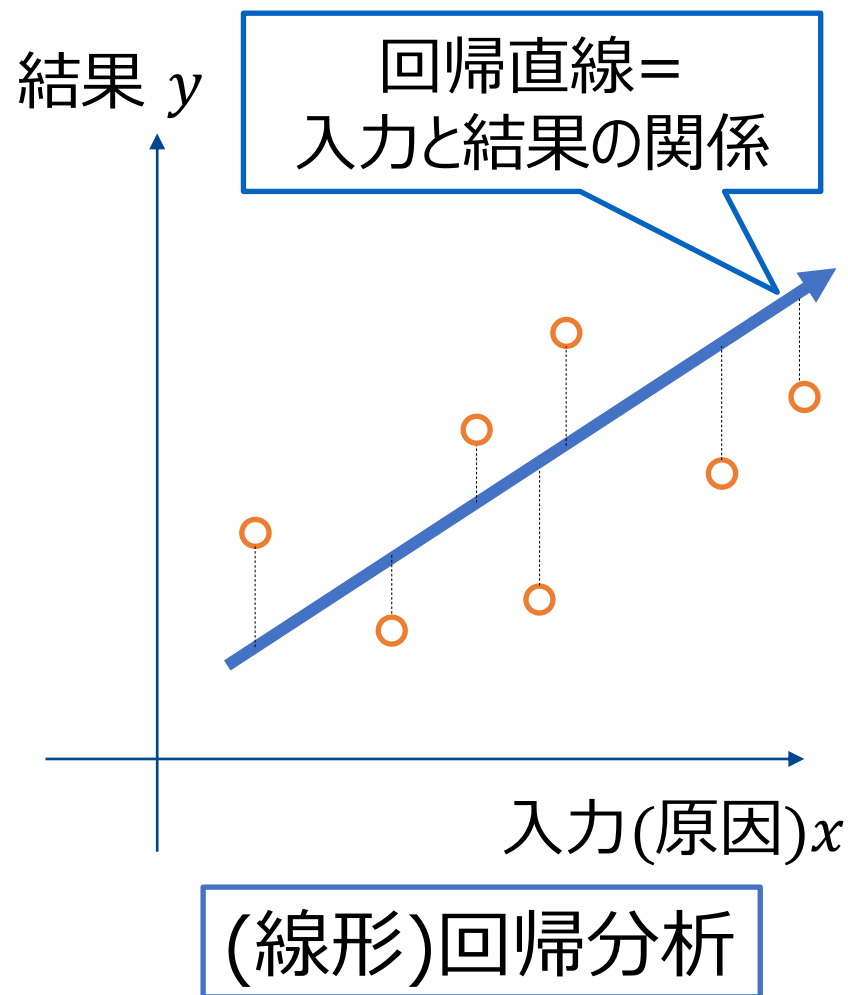


主成分分析と回帰分析の違い：

どちらがいい・悪いというものではなく、目的からして全く違う



要素 x_1, x_2 を交換しても同じ



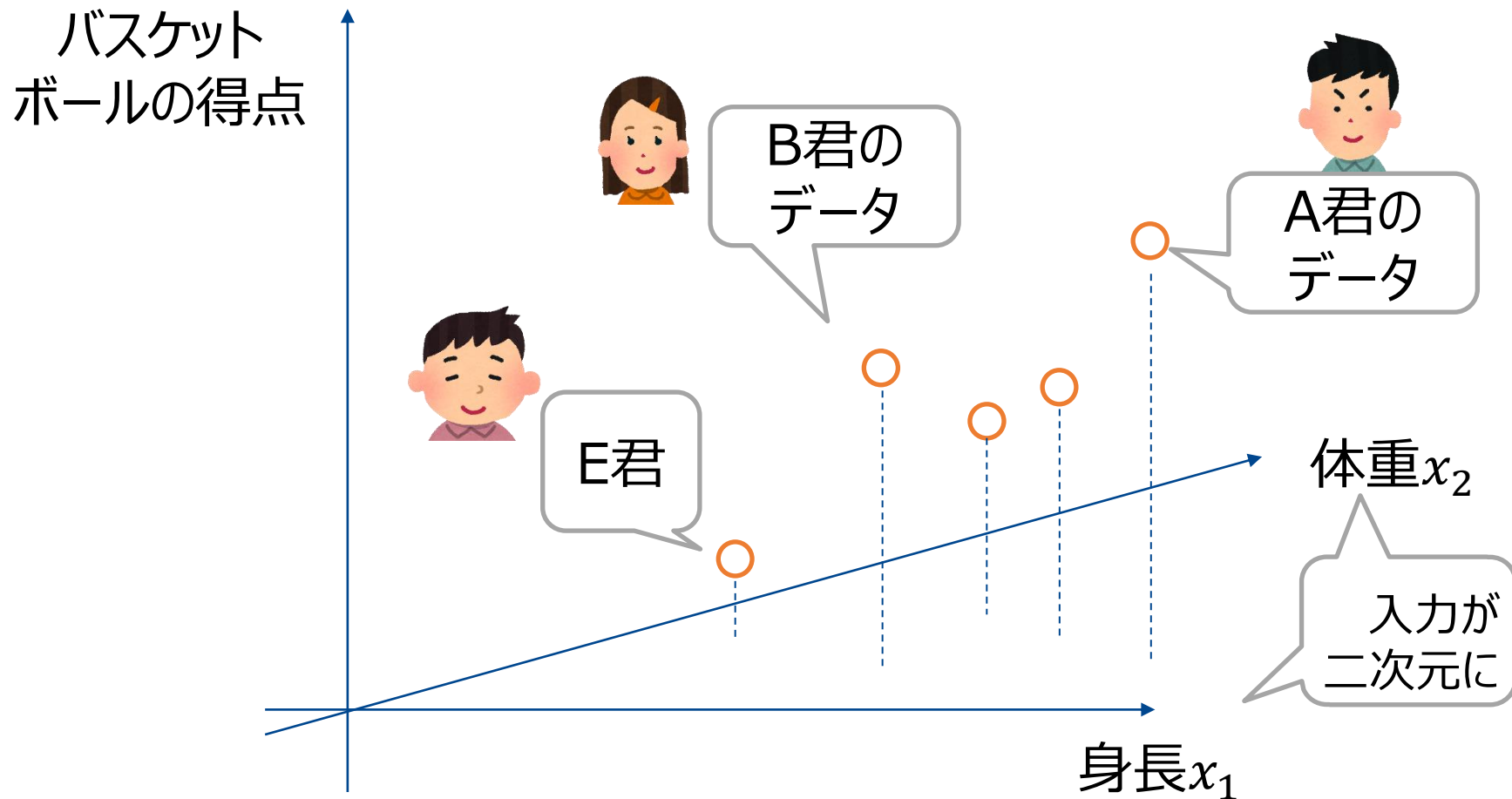
入力 x と結果 y を交換すると違う直線

重回帰分析

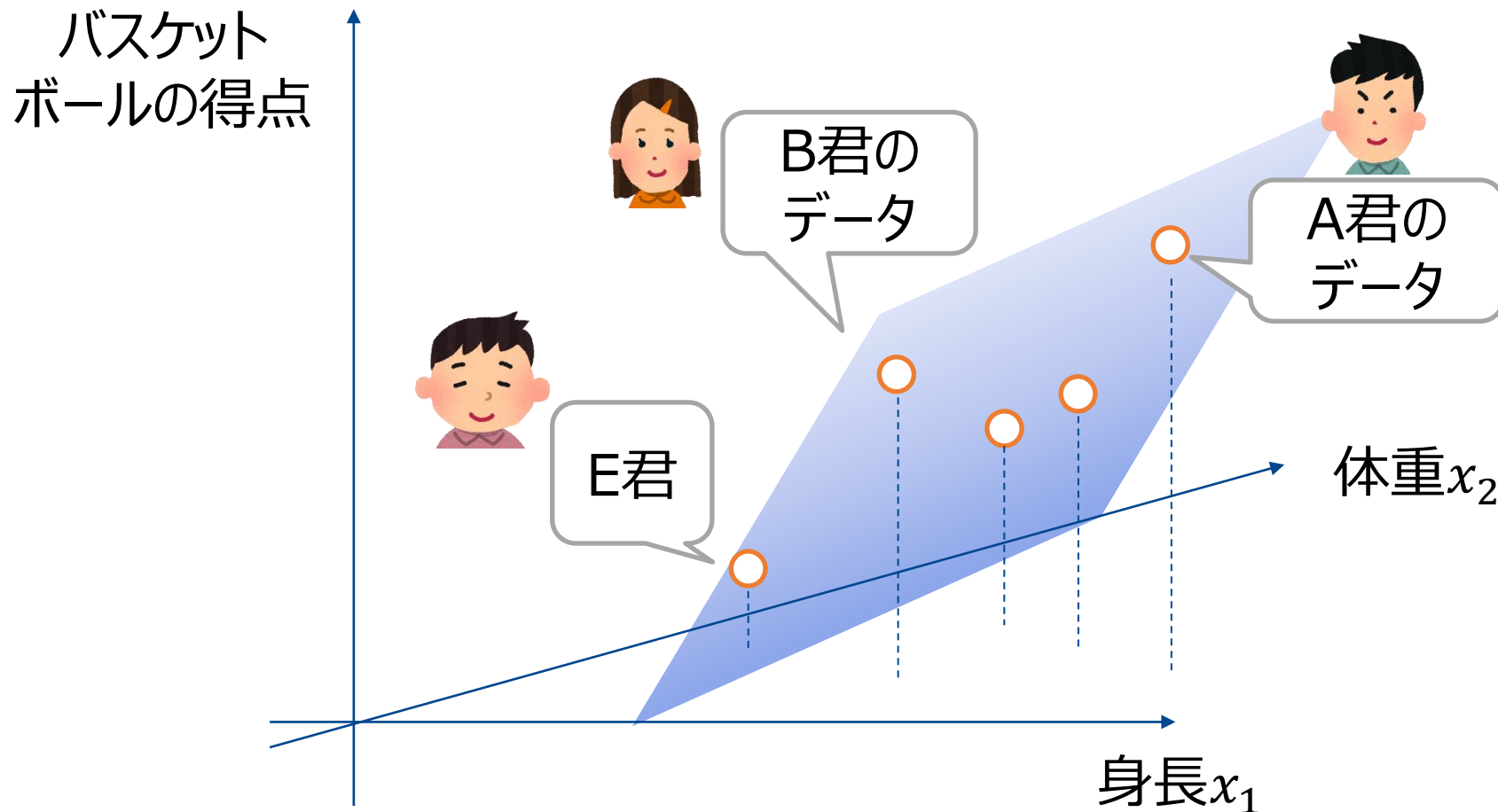
重たくないのに

入力 = 複数の変数
結果 = 単一の値

重回帰による予測（ステップ1/3）： データ収集

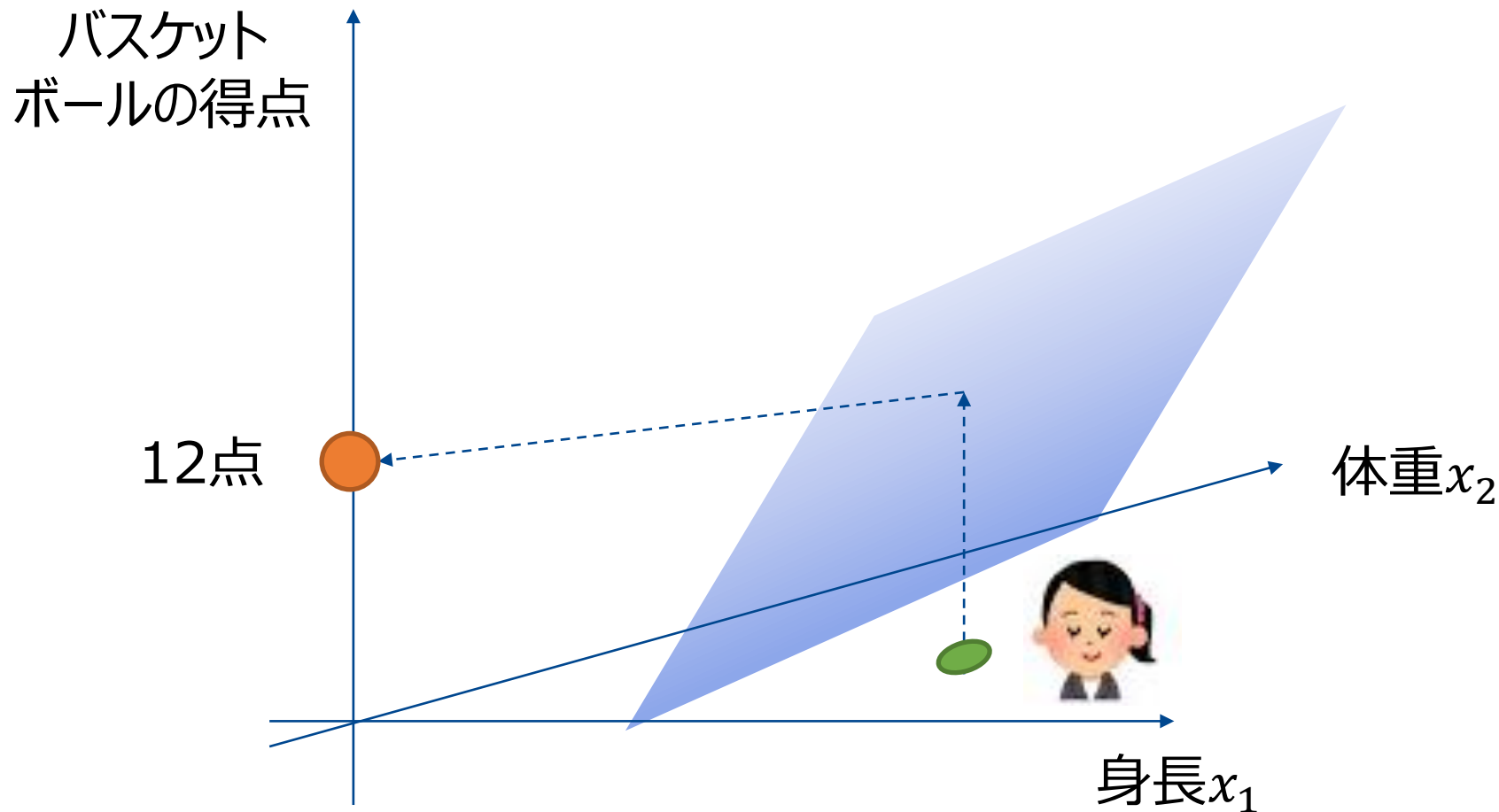


重回帰による予測（ステップ2/3）： モデルあてはめ



→ やはり最小二乗法になります

重回帰による予測（ステップ3/3）： 予測



余談：なんで「重」回帰？ 何か重たいのか？

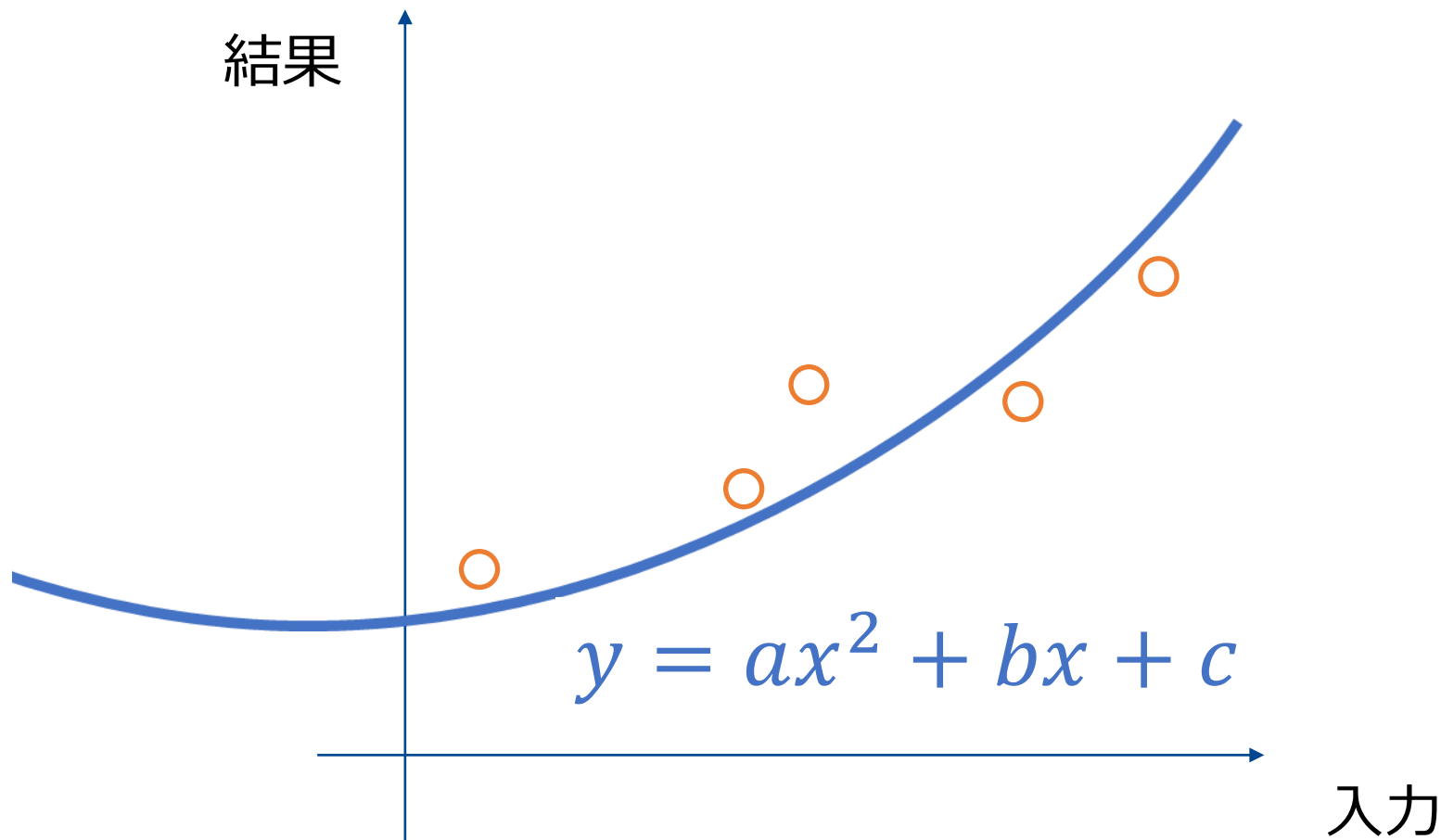
- 英語では **Multiple** regression analysis
 - このmultipleを「重」と訳した
 - Regression analysis = 回帰分析
- Multiple = 「複数のものを(同時に)」という意味合いか
- というわけで「何かが重たい」わけではない



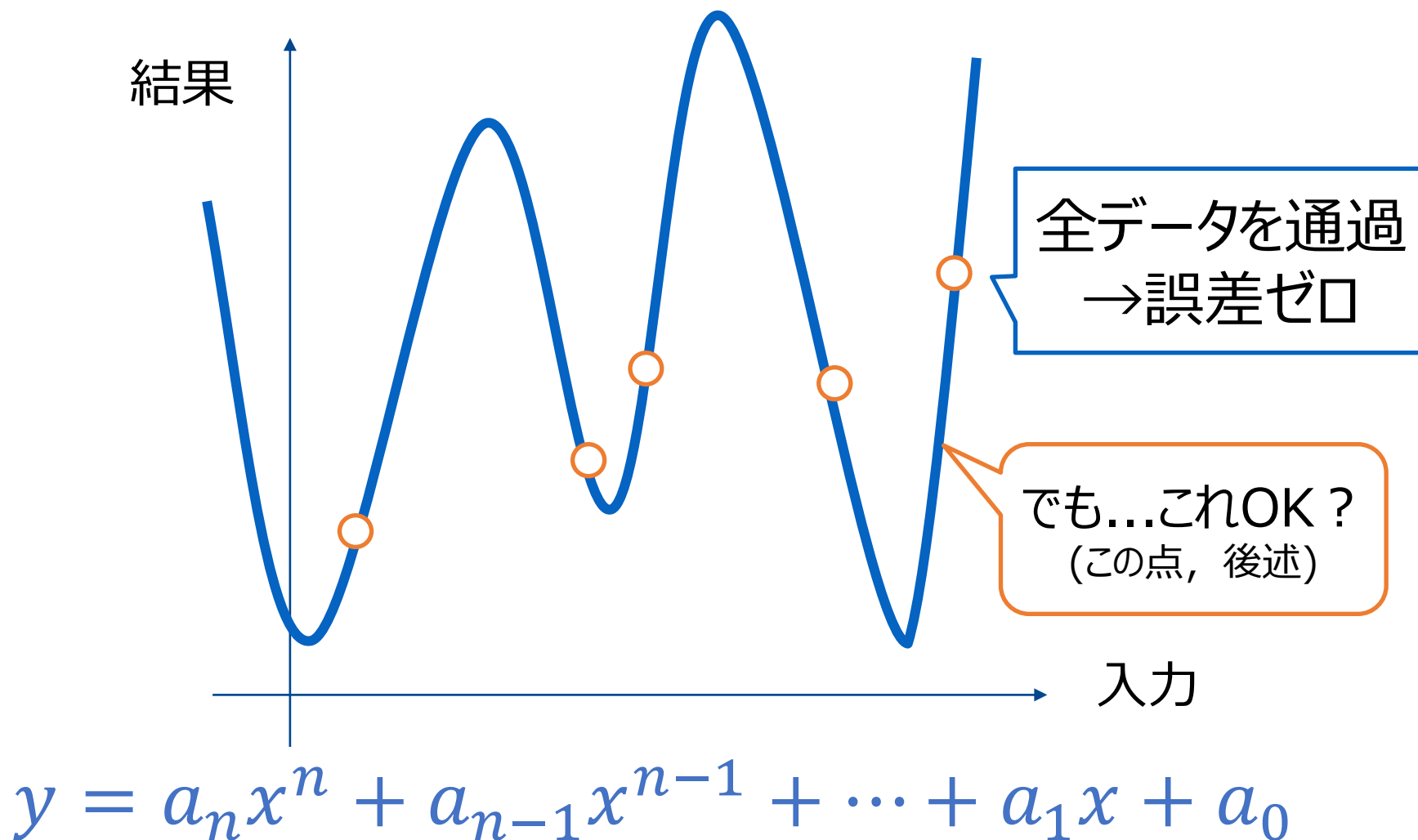
より複雑なモデルの利用

多項式モデル

回帰分析は「線形モデル」だけじゃない： 2次関数モデル



回帰分析は「線形モデル」だけじゃない： N 次多項式モデル



回帰分析で注意したい点

アウトライヤとオーバーフィット

アウトライヤ(はずれ値)の悪影響(1/3)

アウトライヤ

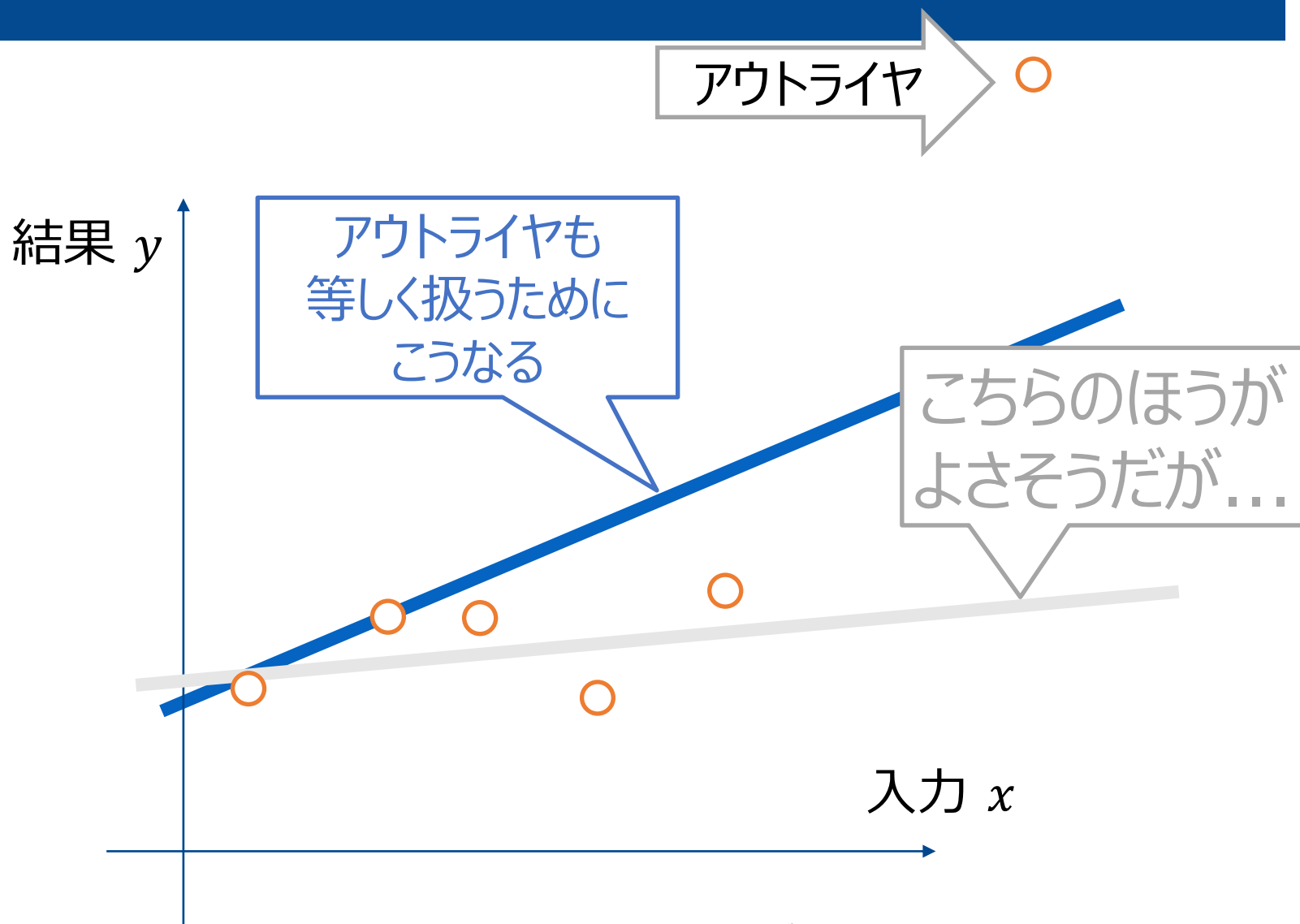


結果 y

- センサの異常
- 測定ミス
- イイカゲンな返答



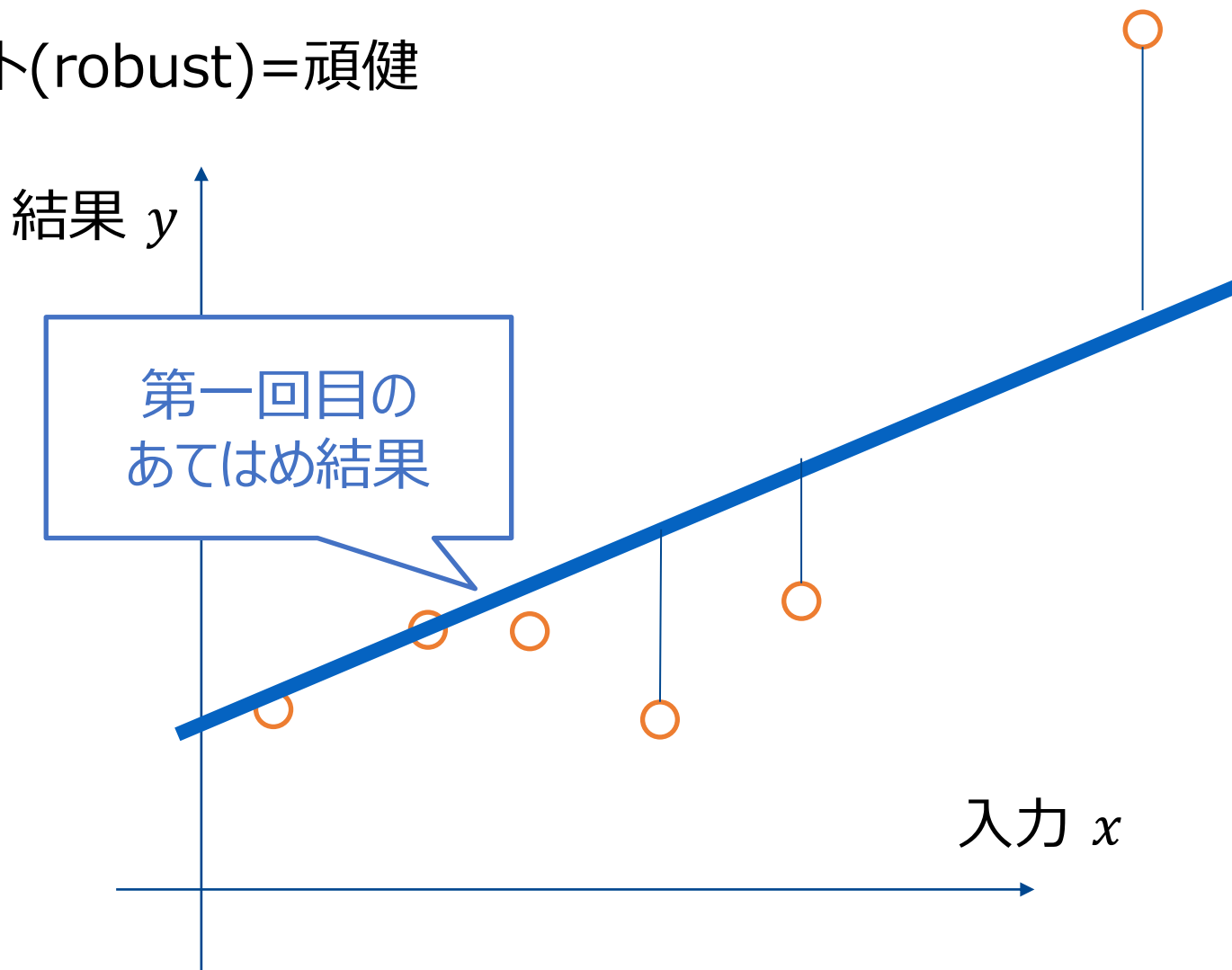
アウトライヤ(はずれ値)の悪影響(2/3)



アウトライヤ(はずれ値)の悪影響(3/3)

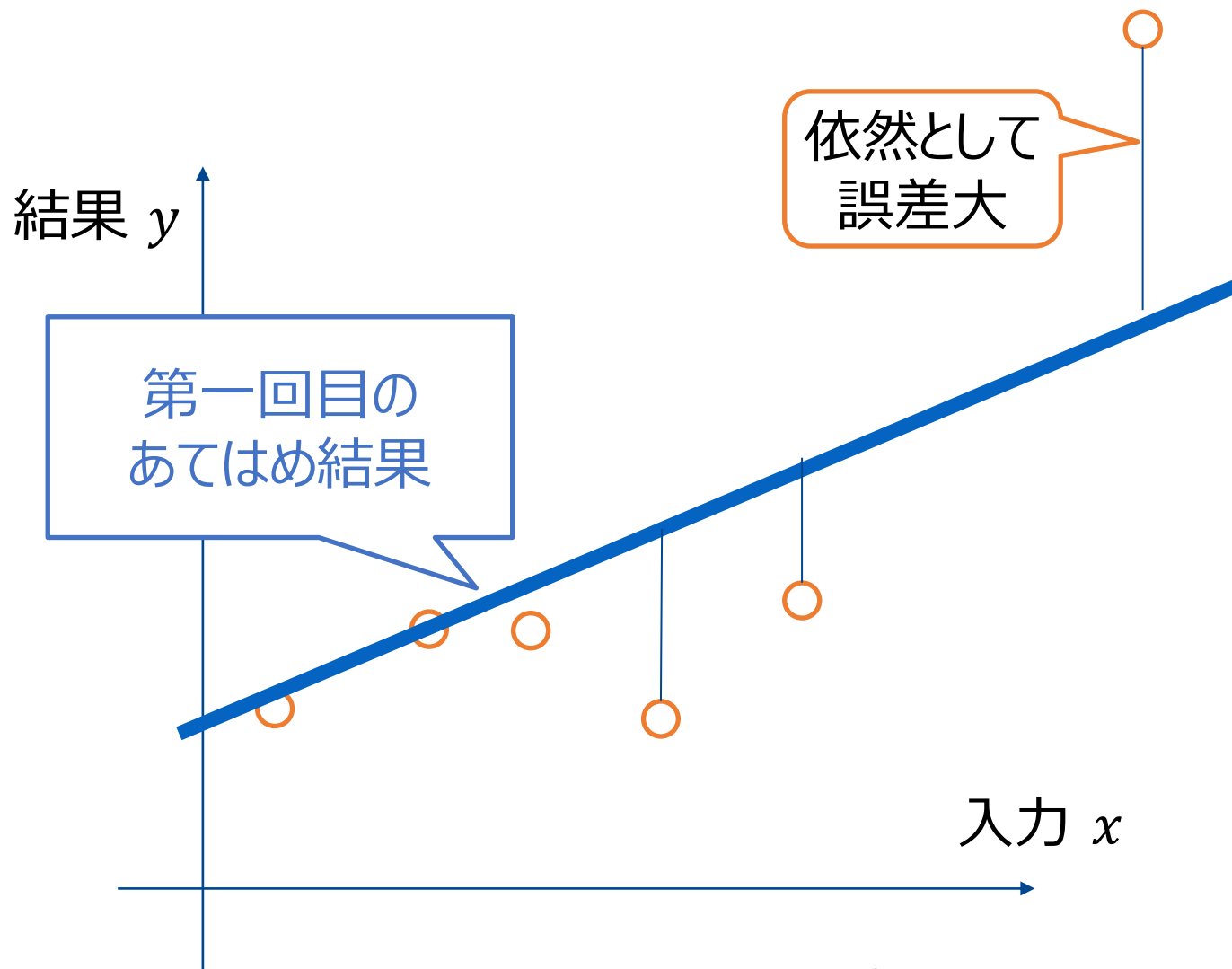
対策:「ロバスト最小二乗法」

- ロバスト(robust)=頑健



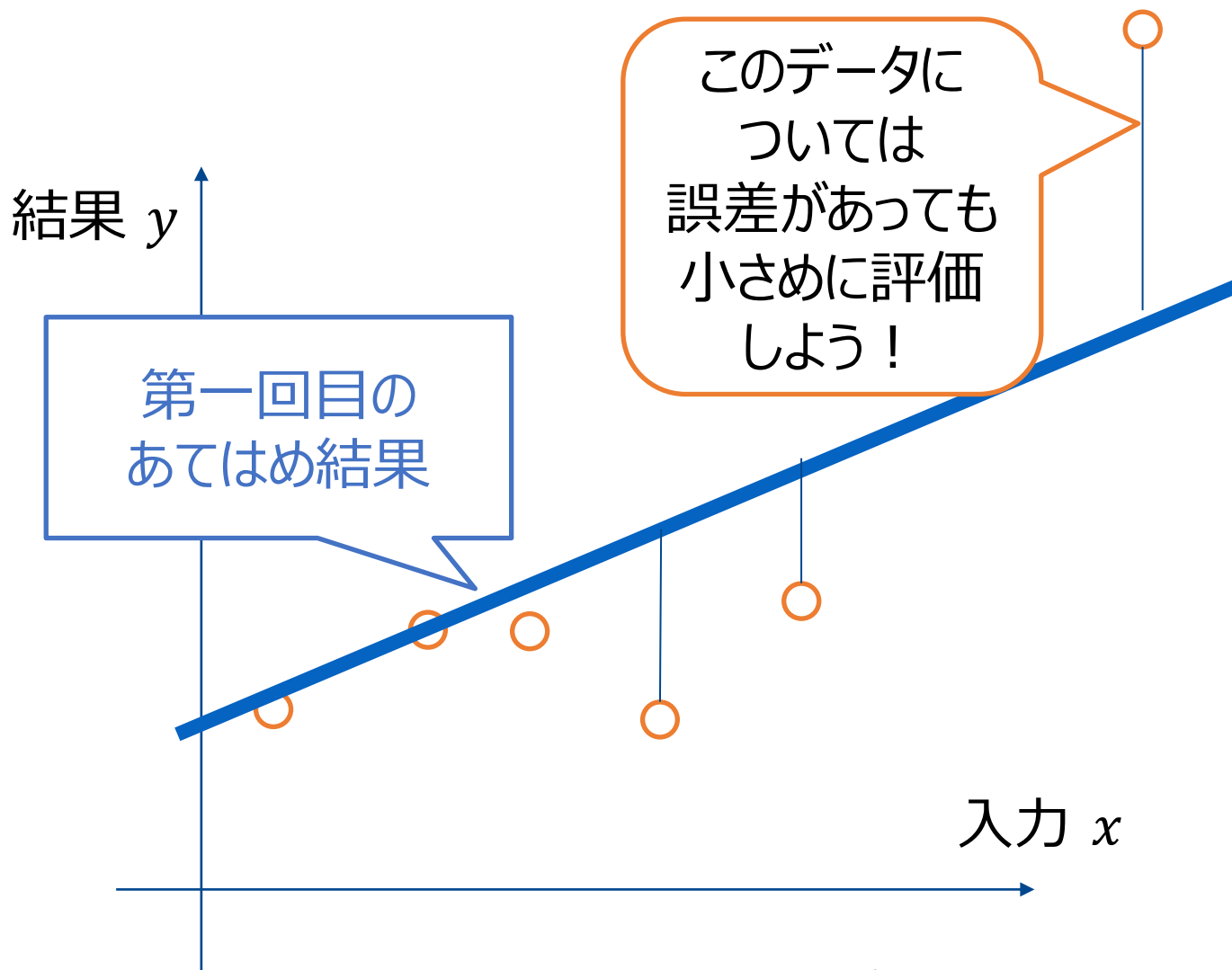
アウトライヤ(はずれ値)の悪影響(3/3)

対策:「ロバスト最小二乗法」



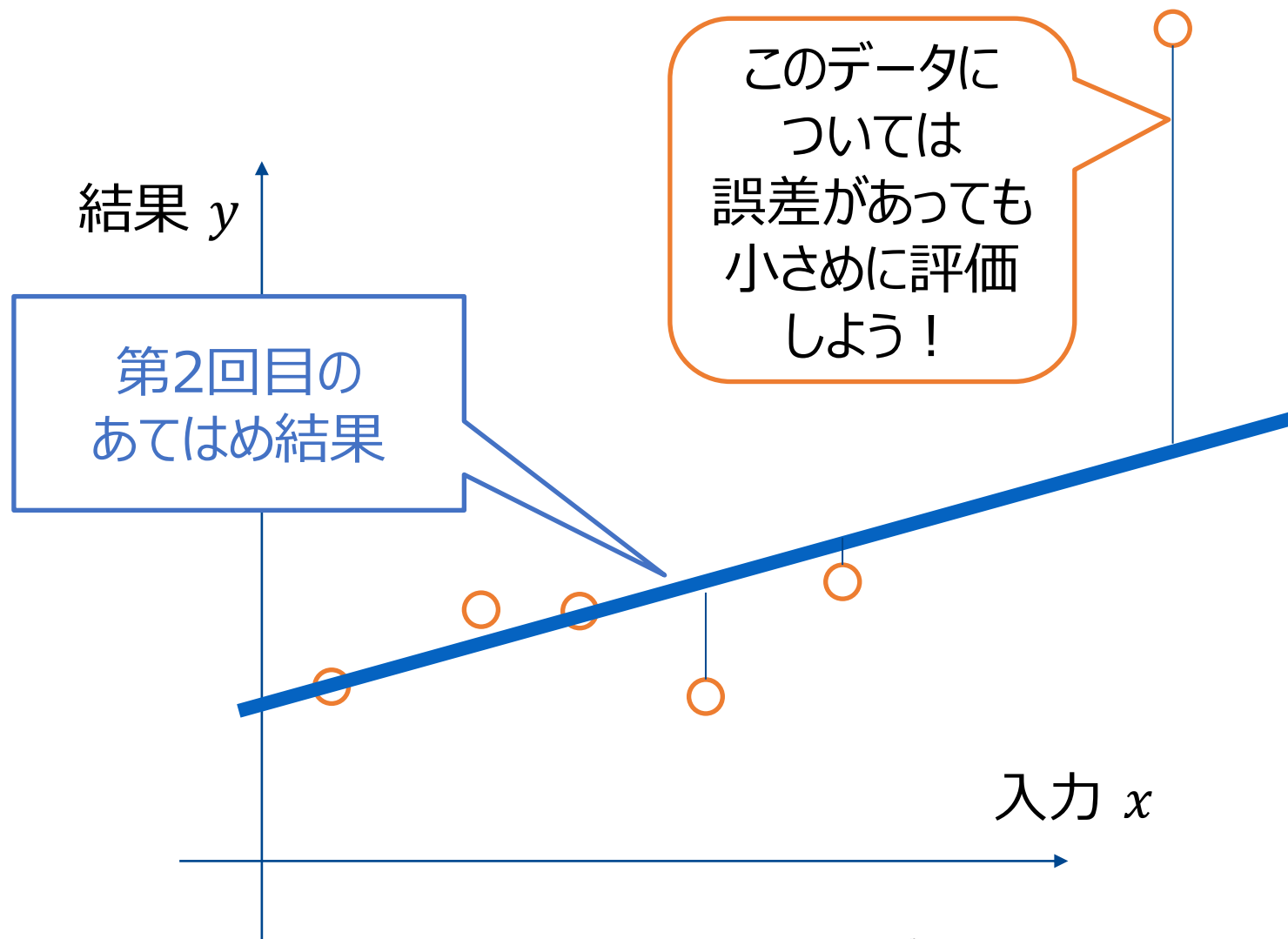
アウトライヤ(はずれ値)の悪影響(3/3)

対策:「ロバスト最小二乗法」

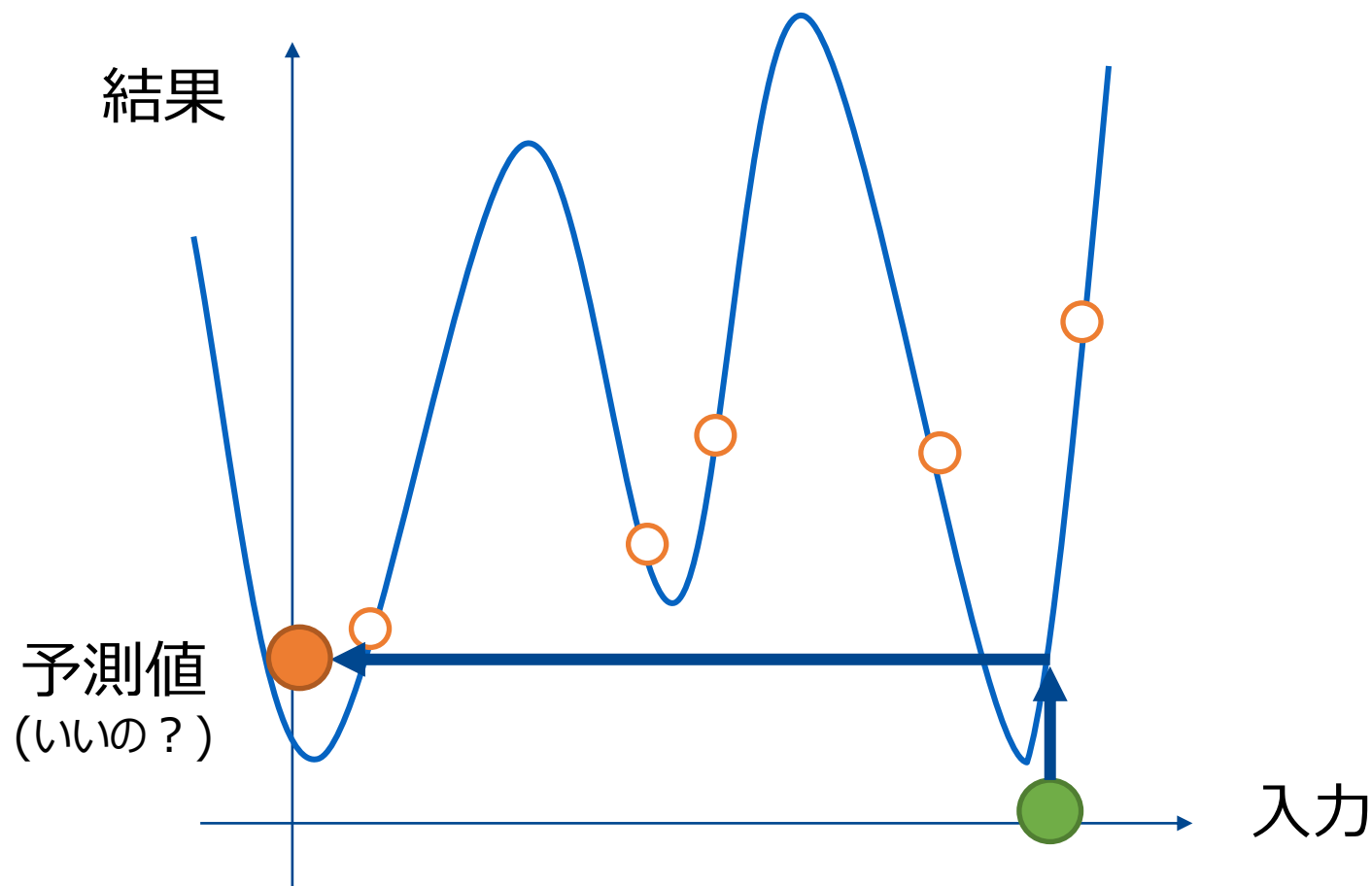


アウトライヤ(はずれ値)の悪影響(3/3)

対策:「ロバスト最小二乗法」



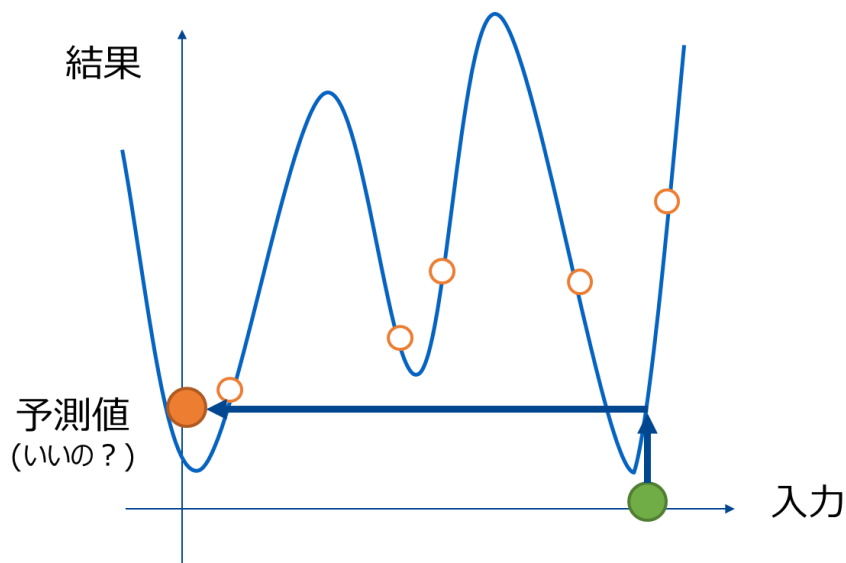
オーバーフィッティングと汎化能力(1/3)



事前に与えられたデータにとってはハッピーだが
未知データにとっては使い物にならない(オーバーフィット) 1月版

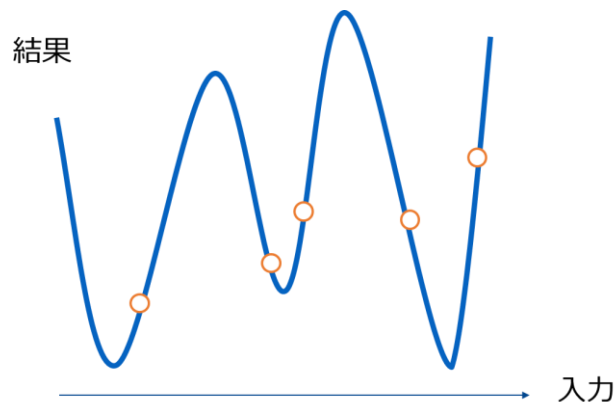
オーバーフィッティングと汎化能力(2/3)

- 汎化能力(generalization)とは？
 - 回帰曲線を求めるときには使わなかったデータについても、ちゃんと妥当な予測結果が得られるか？
- だから下の例は「汎化能力」がない例



オーバーフィッティングと汎化能力(3/3)

高次多項式モデル

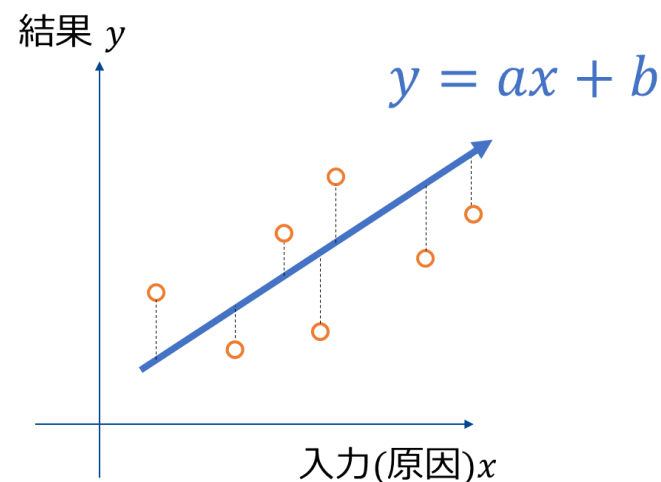


$$y = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0$$

(良) 事前に与えられたデータに対する誤差小さい

(悪) 大量に事前データがないと
回帰結果に汎化能力がない恐れ

線形モデル

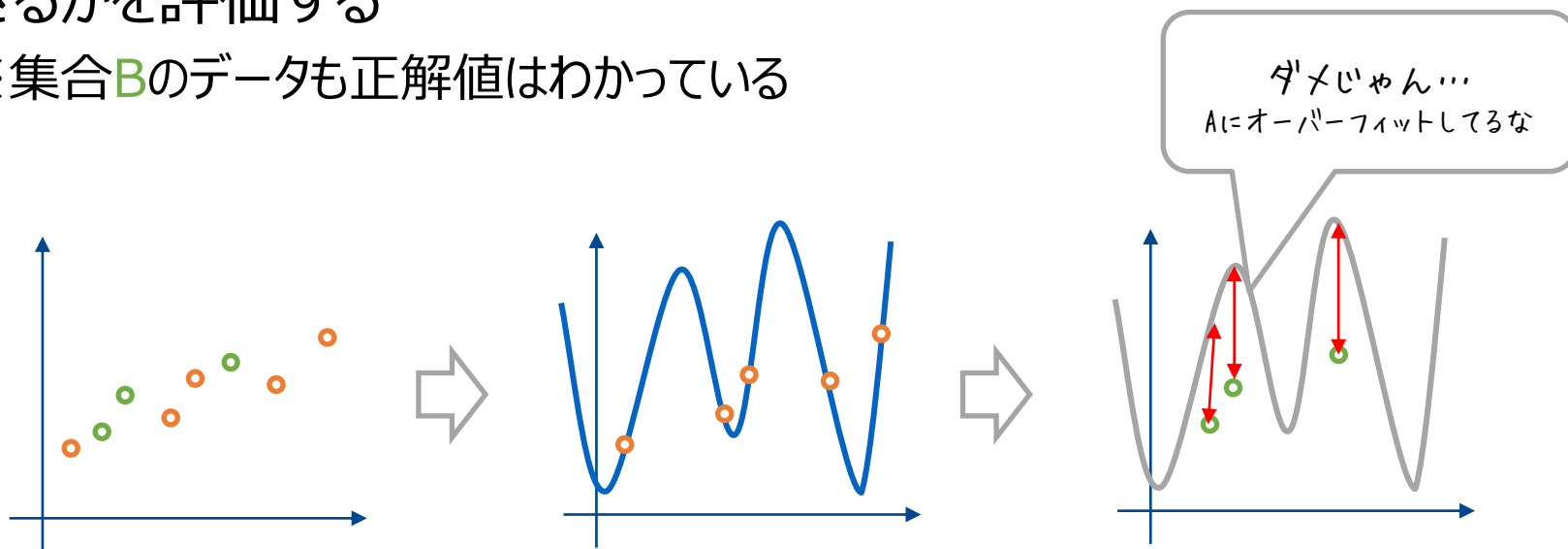


(悪) 事前に与えられたデータに対する誤差大

(良) 事前データ数が少なくても
ひどいオーバーフィッティングは少ない

オーバーフィッティングやアウトライヤの影響を見つける方法：バリデーション(validation)

1. 事前に収集されたデータ集合をランダムにA, Bに分ける
 2. 集合Aの中のデータを使って、モデルフィッティングを行う
 3. 求まったモデルを使い、集合Bのデータがどれぐらいきちんと予測できるかを評価する
- ※集合Bのデータも正解値はわかっている



- 面倒そうですがAI/機械学習では必ず利用します

オーバーフィッティングやアウトライヤの影響を見つける方法：バリデーション(validation)

- 「問題集」で考えてみよう
- 問題集を買って、答えを見つつ全ページで満点取ろうとする
 - 答えを全部暗記してしまえば、全クリア！
 - これって「頭よくなってる」？「わかってる」？
 - この状態がオーバーフィッティング
 - 理解しているわけではないので、初めて見るテスト問題は全く解けない
- 問題集を買って、答えを見つつ2/3で勉強、残り部分でテスト
 - 勉強した部分が満点でも、テスト問題がボロボロならオーバーフィッティング
 - 上のケースと同様、理解せずに暗記しただけ！
 - 勉強した部分が満点でなくても、テスト問題が同程度解けるならOK!

