

一般教育演習

プログラミングで問題を解く：

集計から人工知能まで

瀧川一学

工学部 情報理工学コース

残りの4回！

13. 画像を認識するAIプログラムを作ってみよう(1)

→ (解説) AIの原理を簡単に解説

14. 画像を認識するAIプログラムを作ってみよう(2)

→ (実習) numpyとscikit-learn

15. 授業のまとめと振り返り:この後どうする?

→ (解説) 最初の授業の前フリ(プログラムの役割など)を思い出し
授業でやらなかったこと、今後のプログラミングとの付き合い方、
python以外の言語、などを総括します。

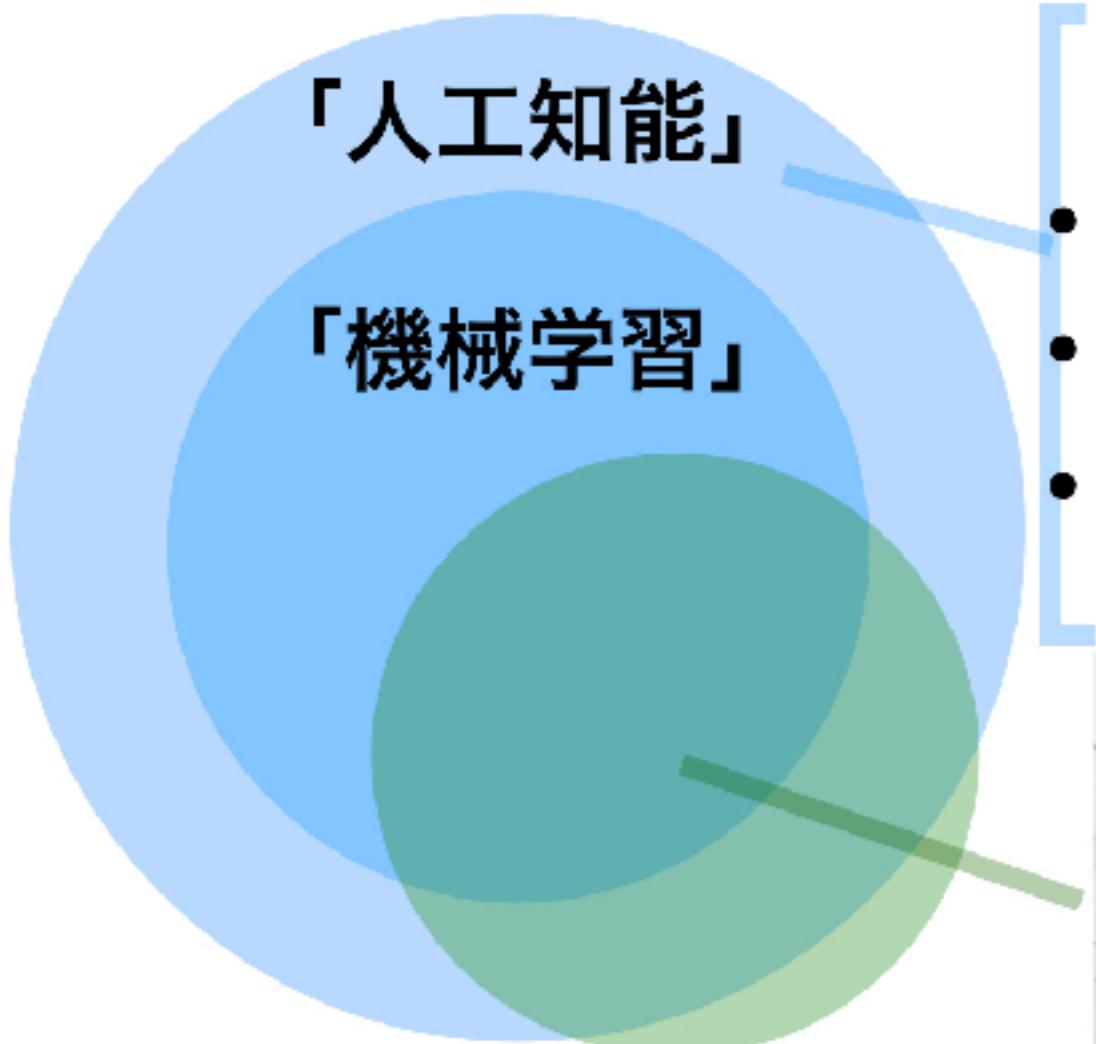
16. まとめNotebook作成日

→ (実習)

今日のお題：Pythonの基本

- フォローアップQ & A
- AIについて
- 機械学習について
- numpyで配列データを扱おう

機械学習は人工知能(AI)実現のための一つの要素技術

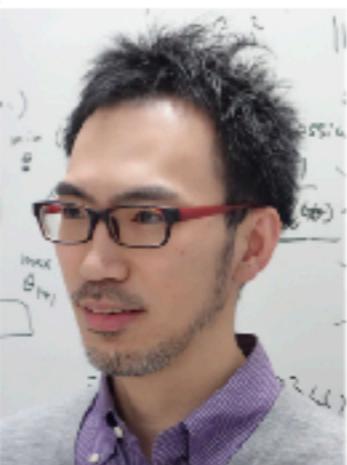


「人工知能」分野 = 複合領域

基礎研究

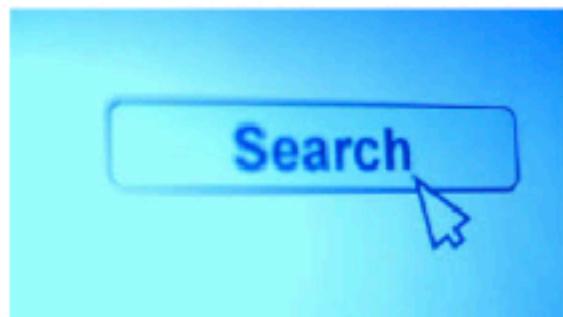
応用研究

- 機械学習
- データベース
- プランニング 制御
- 自然言語処理
- ロボティクス
- 画像認識
- 産業などの利活用
- 離散構造を伴う機械学習
- 科学研究での利活用
(生命・材料科学など)
- データ解析・アルゴリズム



機械学習はさまざまな先端情報技術の基盤です！

検索エンジン



機械翻訳



自動運転



医療診断



広告



物流・需要予測



気象予測



セキュリティ



スマホ、ネット通販、製造業、科学、交通、農業、教育、金融、就職・結婚、…

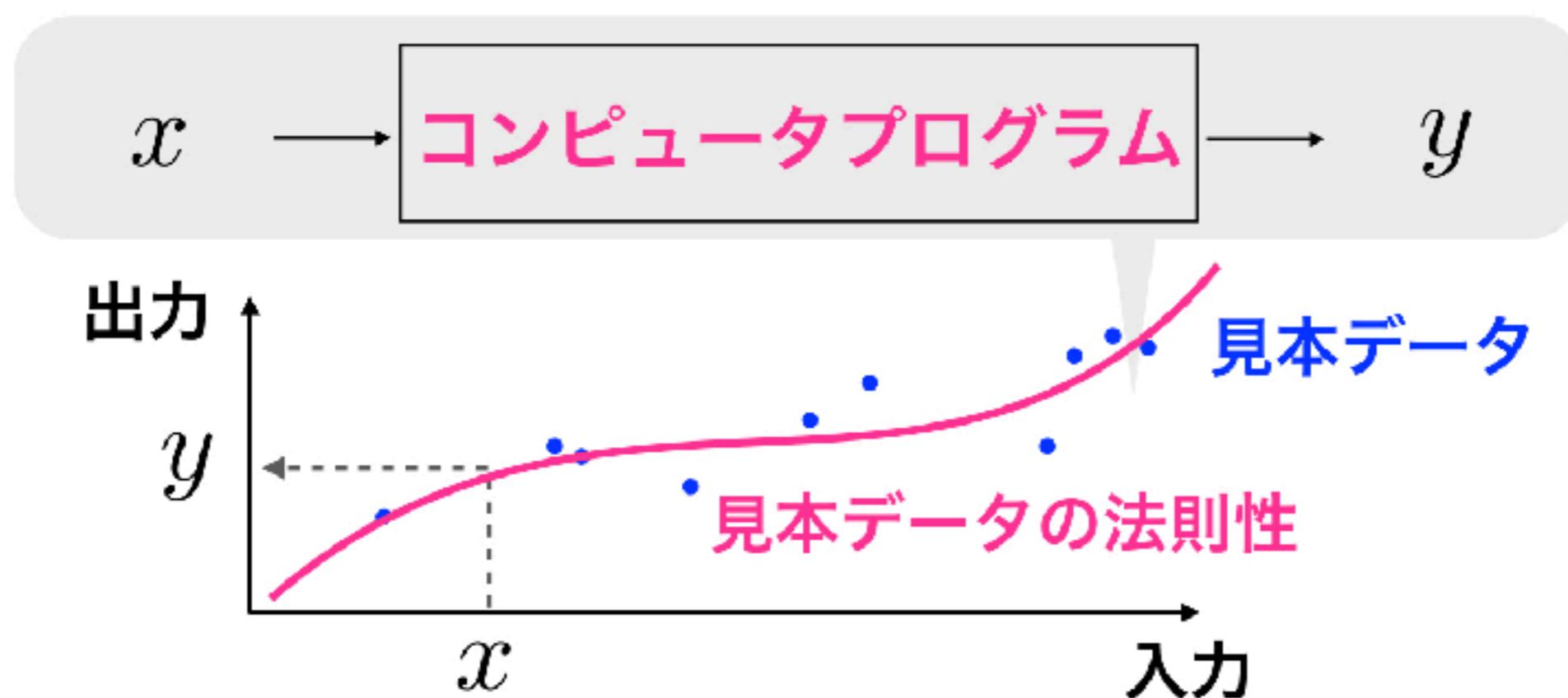
機械学習を説明すると…

多量の入出力の見本からコンピュータプログラムを自動的に作る技術



「機械がデータから学習する」しくみ

機械学習は見本のデータを補間することで結果を予測する



機械学習自体の基礎研究

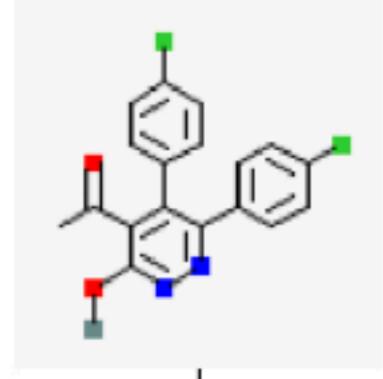
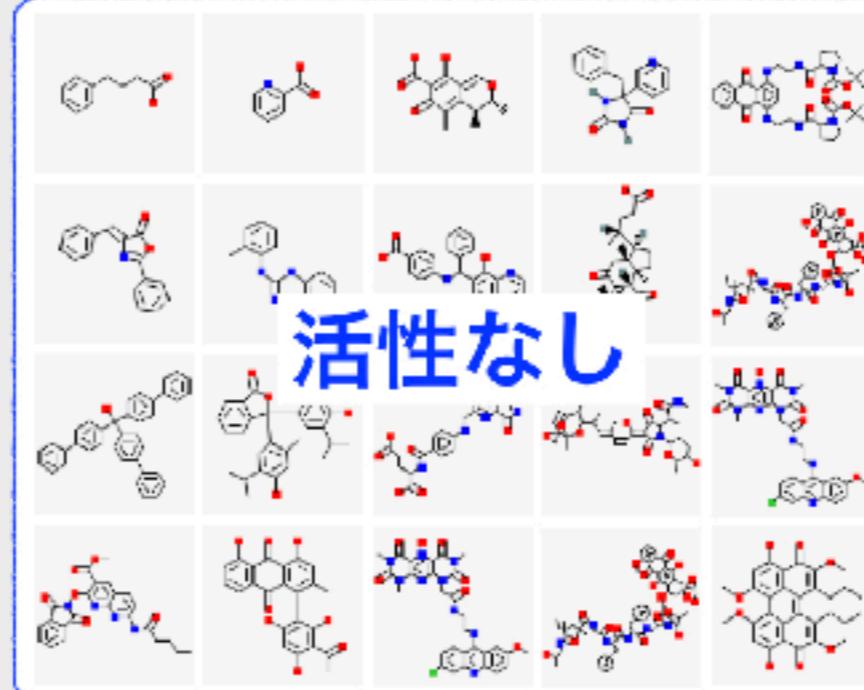
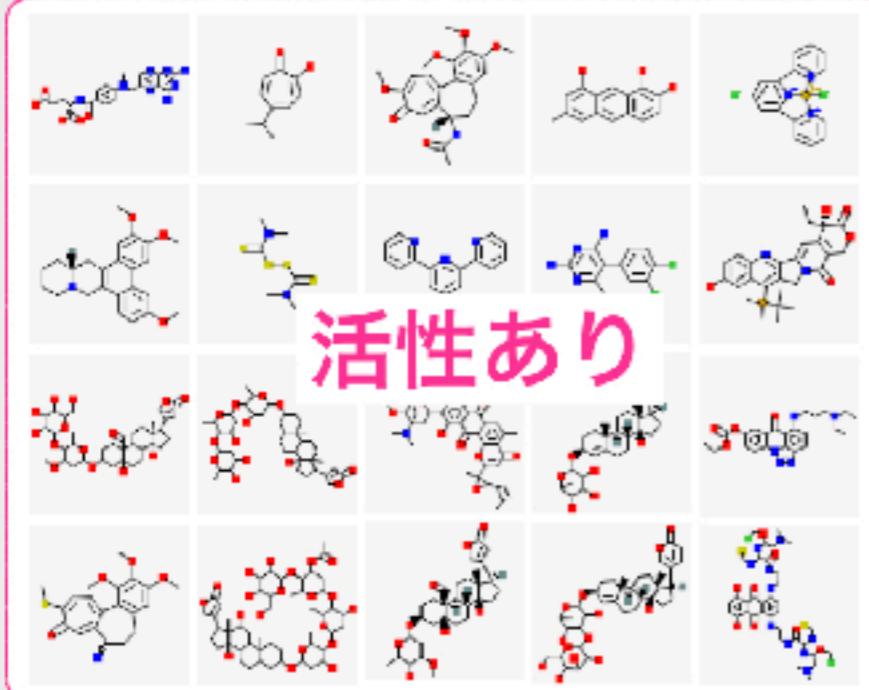
まだまだ発展途上の技術であり**基礎研究**が非常に重要！

- 補間を正確に行う**新しい計算方式**を研究する
- 既に知られた計算方式の**改良や拡張**
- 既に知られた計算方式の**分析、他の知見との関係づけ**

事例：「かたち(数値ではないデータ)」の機械学習

抗がん剤の候補選別：がん細胞の増殖を防ぐ効果があるか？

既に実験済みのデータ(見本)



機械学習予測

活性
あり or なし

質問その1

以下の7つのGoogleサービスを今までに
一つも使ったことがない人はいますか？

- Google検索
- Gmail
- Android
- Google Maps
- YouTube
- Google Chrome
- Google Play

質問その2

以下の7つのGoogleサービスにい今までに
一回もお金を払ったことがない人はいますか？

- Google検索
- Gmail
- Android
- Google Maps
- YouTube
- Google Chrome
- Google Play

なぜ無料で使えるものが多いのか

- ・ この7つは**月間**の実利用者数が**10億**を超える
- ・ 維持だけでも莫大なお金がかかるはず

Google社 (アルファベット社)

2015年の売上高：749億8900ドル（約9兆円）

うち、**広告売上**高：673億9000ドル（約8兆1000億円）



機械学習のニーズ：すべてがデータになる社会

インターネット・スマートデバイス・IoTが関与する業態

- ・ 対象の詳細なデータを大規模に収集することが可能
- ・ 得られるデータを有効利用しなくてはならない
→ 広告/商品推薦、マーケティング、消費者行動理解

「情報技術 (分析担当者の仕事の自動化)」の必要性

- ・ 効率化：そもそも人手でさばける規模を超えている
- ・ 合理化：“マイサイド・バイアス (確証バイアス)”
人はデータに自分の見たいものを見てしまう

広告、マーケティング、消費者行動理解...?

‘The best minds of my generation are thinking about how to **make people click ads**... That sucks.’

Jeff Hammerbacher

真理。しかし、新しい技術はまず**最も手軽に稼げるところ**に導入されるもの。最初に印刷技術で印刷されたものは「聖書・宗教冊子」「政治論説」「ポルノ」

機械学習を始めとする情報利活用技術は**かなり汎用的**なものであり、まだまだ広い適用可能性を持つ！

データ大氾濫と情報過多

- とにかく身の回りでいろいろな情報がデータになり
もはや人手では利活用できない... 😞
 - “Data Deluge” (データの大氾濫)
 - “Information Overload” (情報過多)
- このトレンドは科学研究の世界にも到来 😱
 - 計測機器の高精度・高効率・大規模化
 - 生命科学分野などのオープンデータベース
 - インフォマティクスの重要性の増大

機械学習とは？

「機械が経験から学習する」機能を実現する情報技術

機械 = コンピュータプログラム

経験 = データ

学習 = いくつかタイプがある

- ・ 教師あり学習：お手本から学習
- ・ 教師なし学習：データ自身のパターンを学習
- ・ 強化学習：試行錯誤で学習

もう少し詳しく... 「学習」 のタイプ

- ・教師あり学習 (**supervised**)

入力と出力のたくさんの見本から入力→出力の計算を学習

- ・教師なし学習 (**unsupervised**)

与えられたデータのみで何とかする学習

例) 似ているものをグループ化, 次元削減, 埋め込み, 多様体学習

- ・半教師あり学習 (**semisupervised**)

上二つの間くらいで色々設定がある

(教師なしのデータが混じってる、グループ化に見本ついてる、etc)

- ・強化学習 (**reinforcement**)

試行錯誤してみて成功失敗のフィードバックで学習

例) 赤ちゃんの歩行の学習、運動やゲームプレイの上達

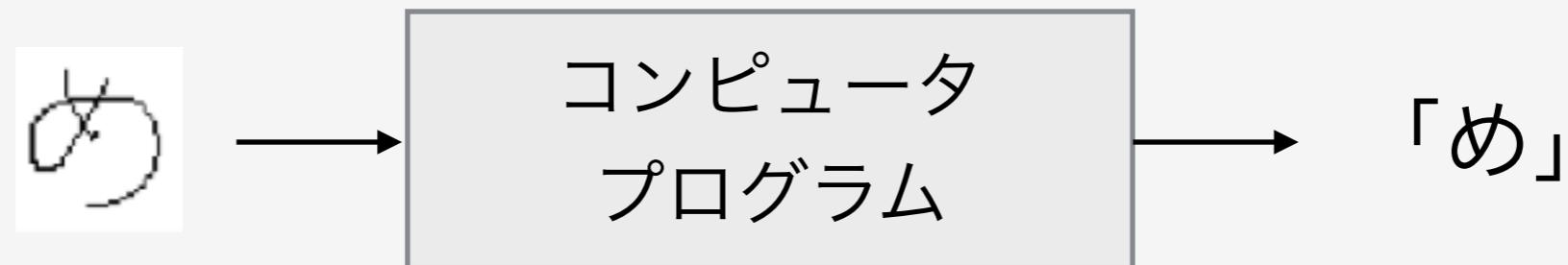
→今日はこれを説明

実問題を生きるために：論理推論から統計へ

- ・ その昔「知能」とは単純に「論理的思考」を意味した
- ・ 人間が「知能」がないとできないと思っていたタスクの多くは思ったほど論理的ではないことが判明
- ・ すべてを論理で厳密に推論(or 形式化)しようとするとすぐ組合せ爆発やフレームや暗黙知(常識)や自己言及などの問題が生じて実問題が全然解けない
- ・ 蓄積されたデータに潜む法則性を経験則として取り出して、それをフル活用するほうが実問題が解ける！

事例：手書き文字認識

入力した手書き文字の画像を入力すると、その文字を出力してくれるようなコンピュータプログラムをつくりたい



...しかし、どういうコードを書けばいいのだろう?

人間にはたやすくできる処理だけど、そもそも我々自身どうやってこの作業をやっているのか理路整然とは説明できん...

コンピュータプログラムをつくるとは

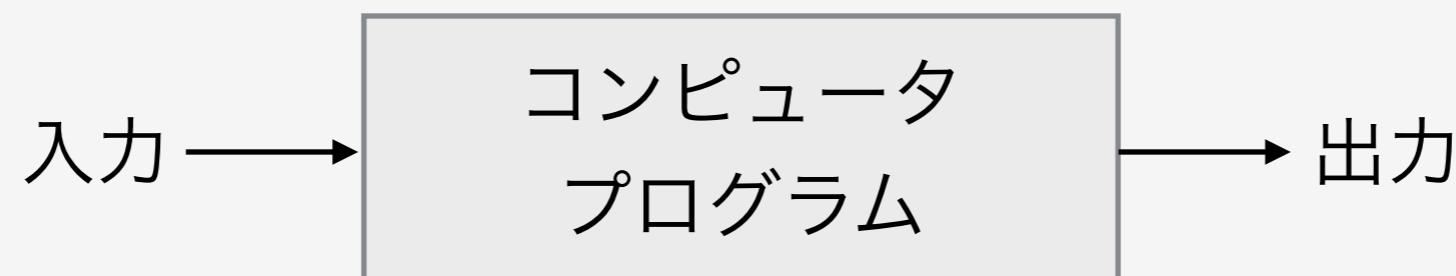


会社の給料計算も、銀行の口座管理も、ミサイルの弾道計算も、スマホアプリも、プレステのゲームも

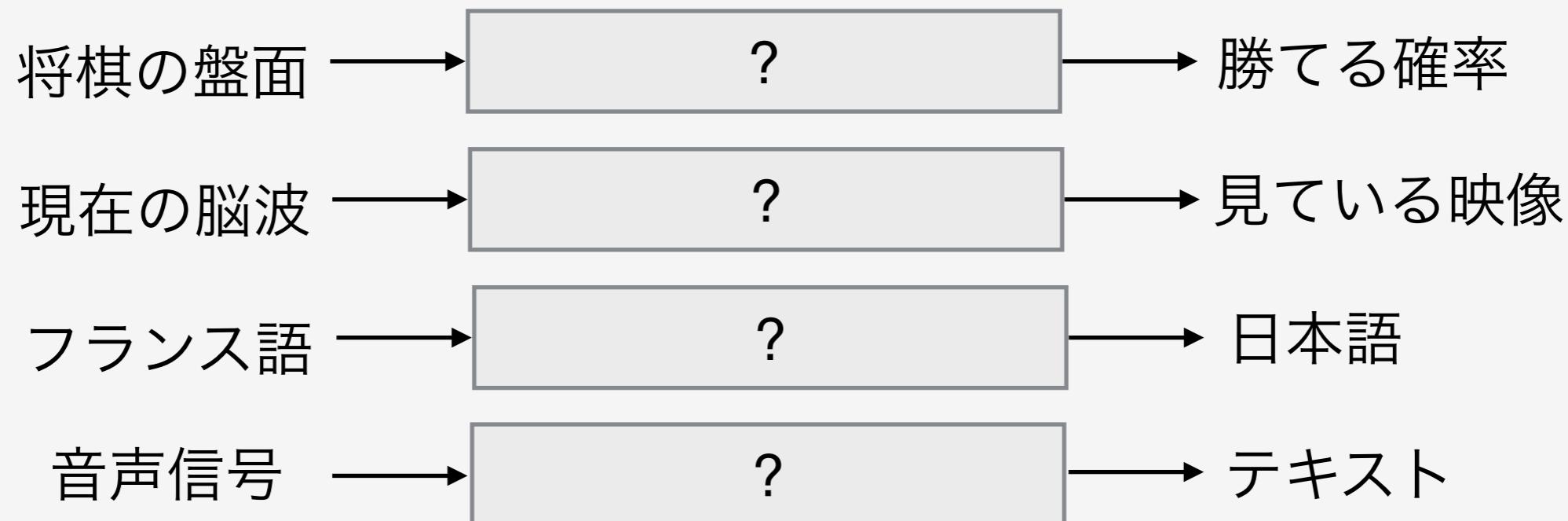
「計算」部のロジックは人間が考え、通常はプログラミング言語を用いてがしがし記述する

教師つき学習：お手本からの学習

コンピュータプログラムを作るには入力から出力を得る手順が**カンペキ**にわかってなければいけない。

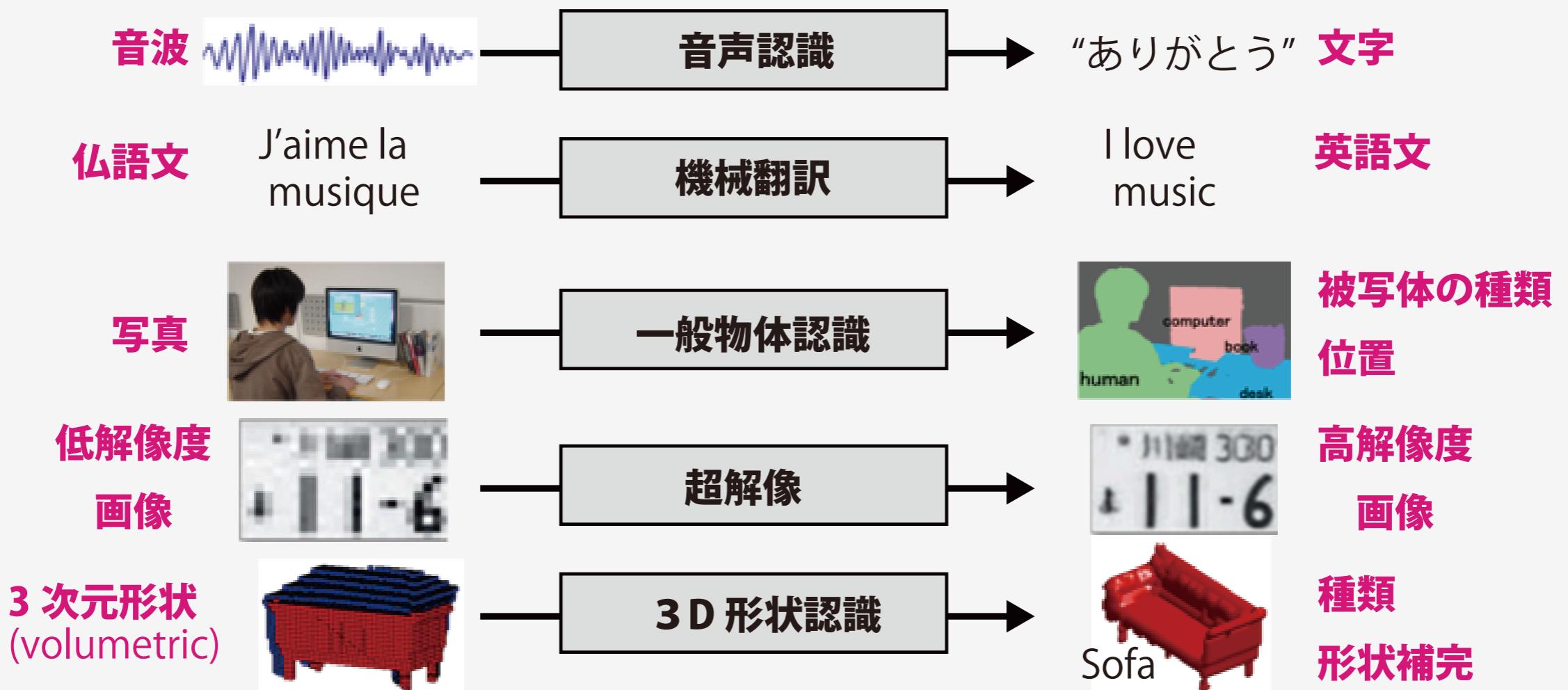


でも、実際には**肝心の手順がよく分からぬことが多い**



教師つき学習 = 新しいプログラミングの形

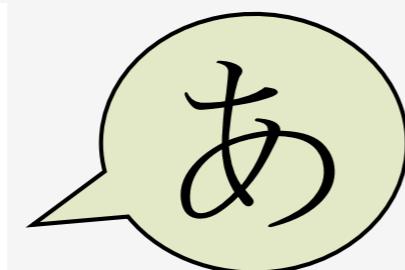
教師つき学習 = 入出力の多量の見本から計算部
のロジックを自動的に構築する技術



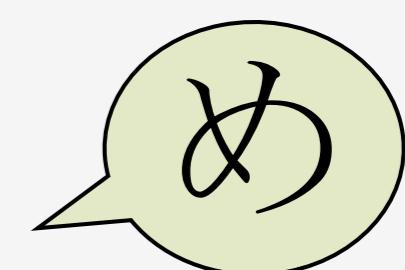
教師つき学習のしくみ：統計的法則性の活用

一つの事例だけ見てもわからないが多くの事例に共通して浮かび上がってくる統計的な法則性を使う

あ あ あ あ あ
あ あ あ あ あ



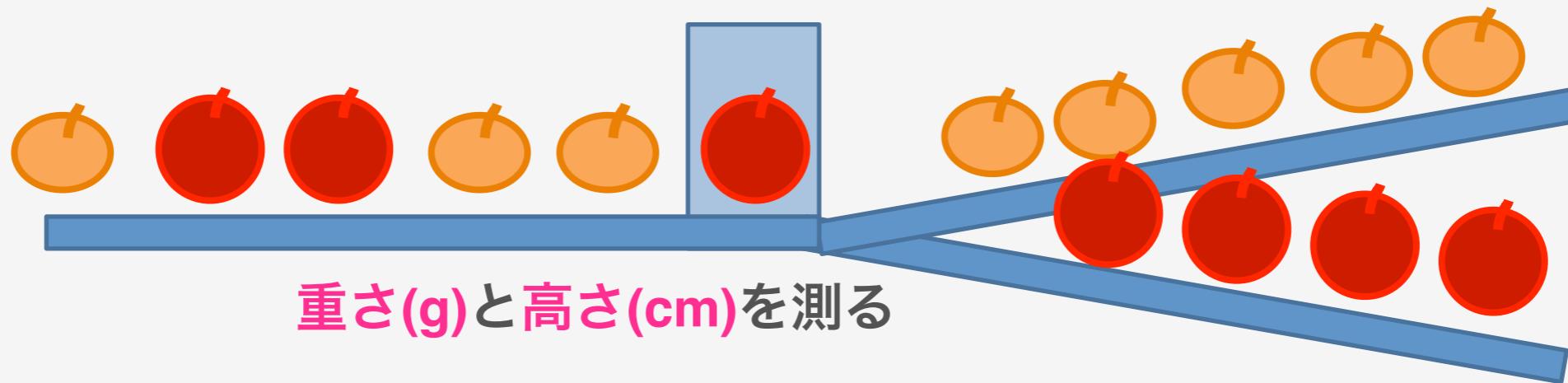
め め め め め
め め め め め



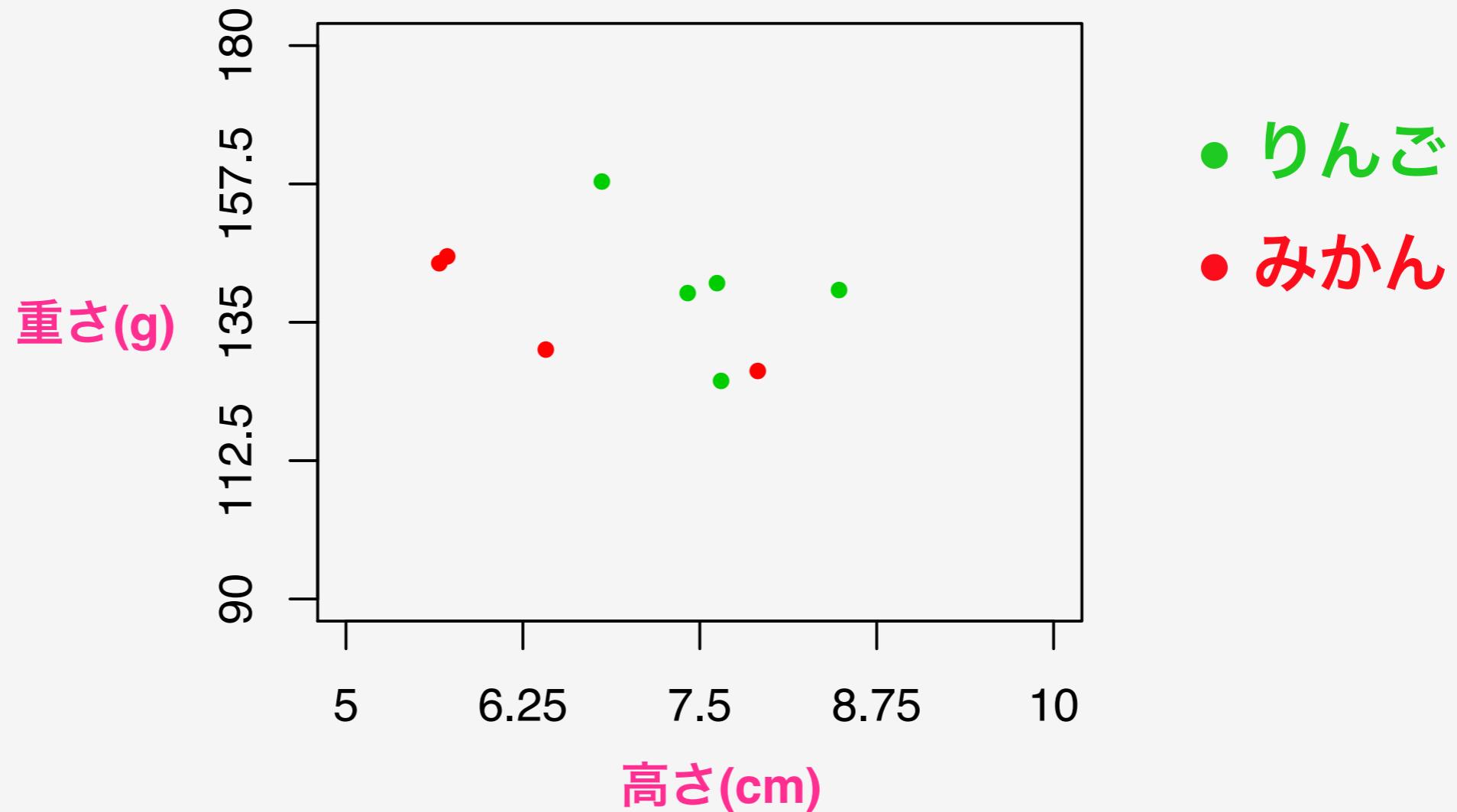
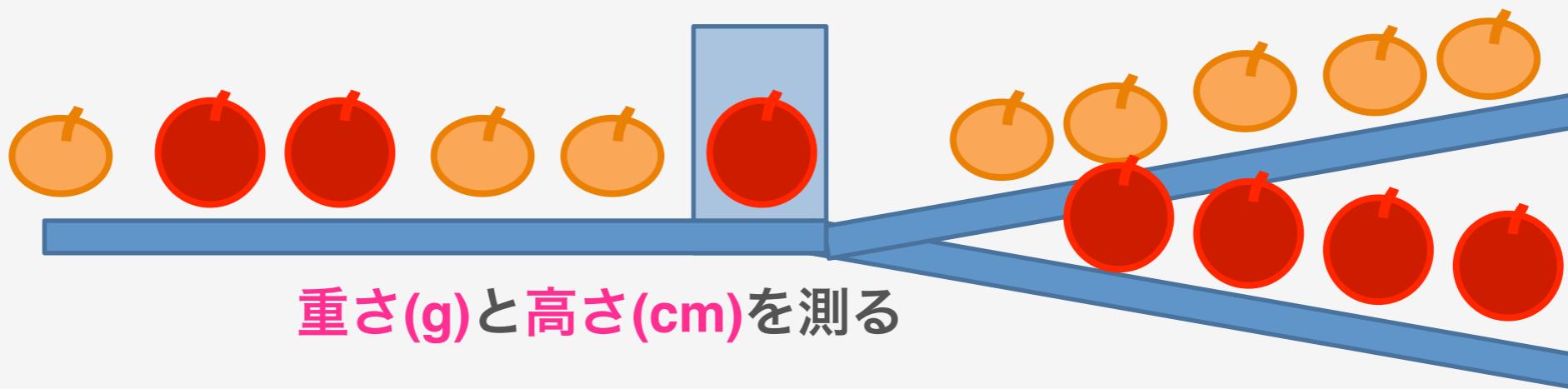
(現在の)機械学習 = データの抽象化 = データ中の統計的法則性を機械(コンピュータ)に学習させる理論

他の例) スマホのカメラかざしたら人の顔がうつっている箇所に枠が表示される機能をどうプログラムする？

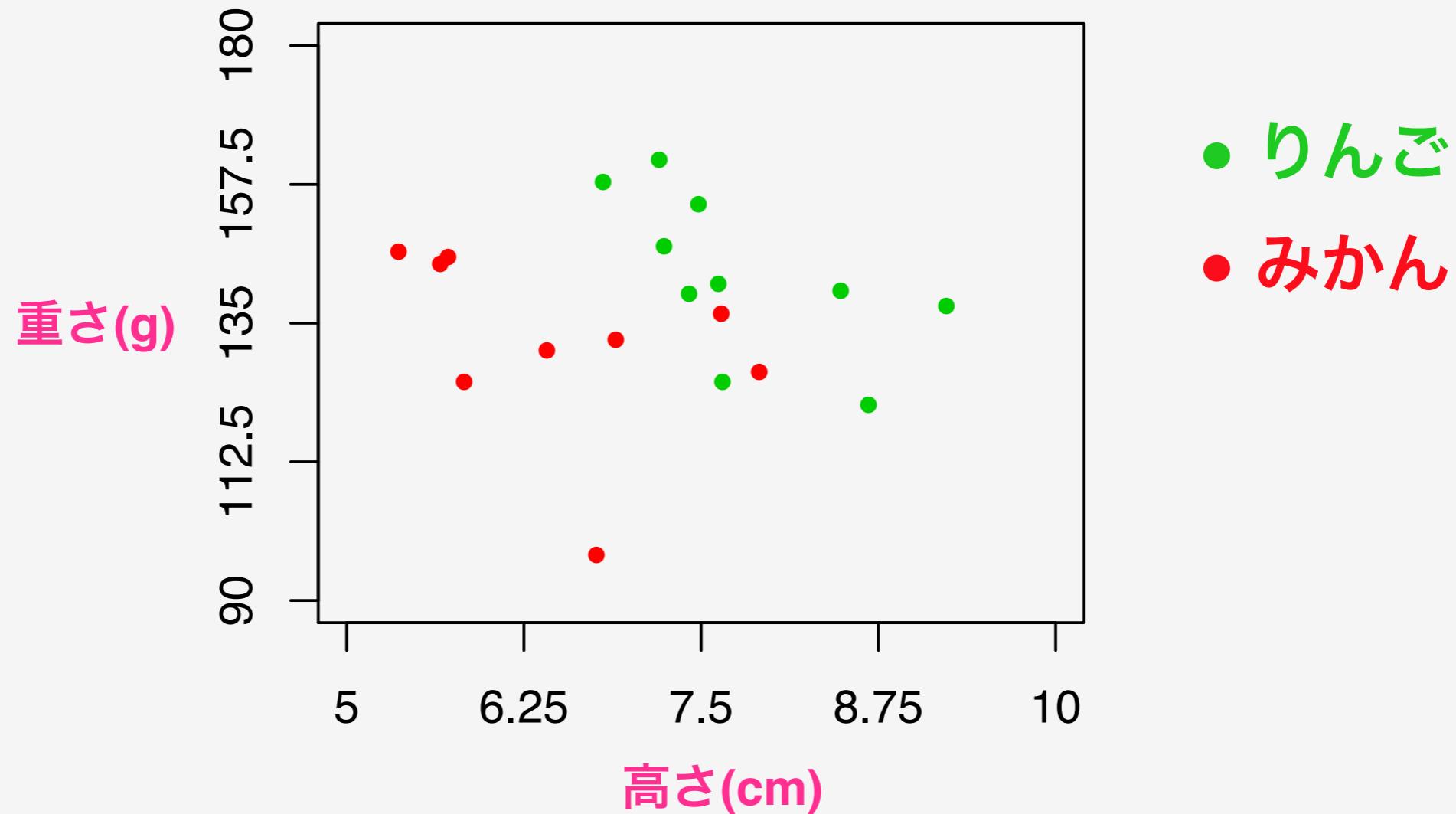
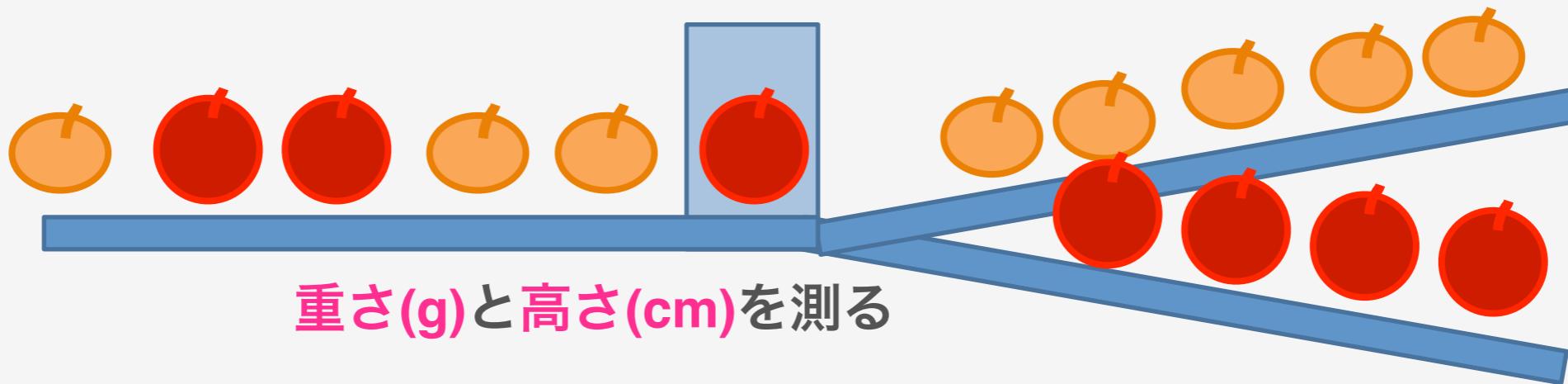
統計的に浮かび上がる法則性



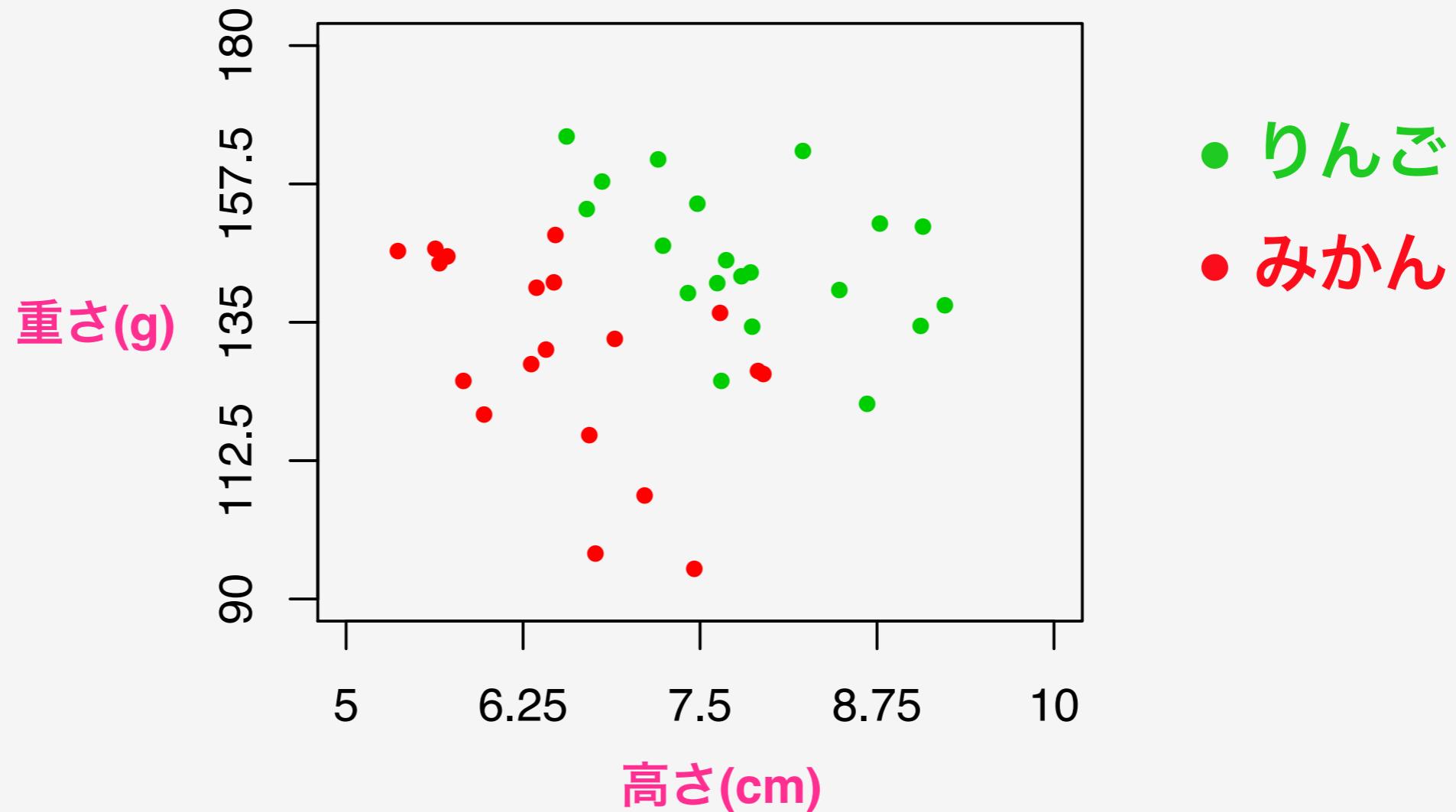
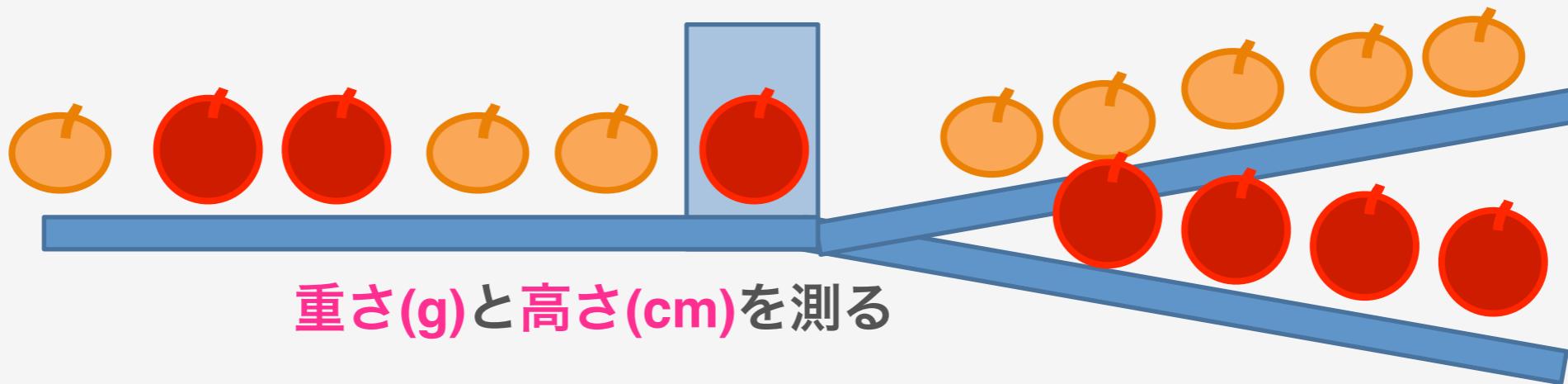
統計的に浮かび上がる法則性



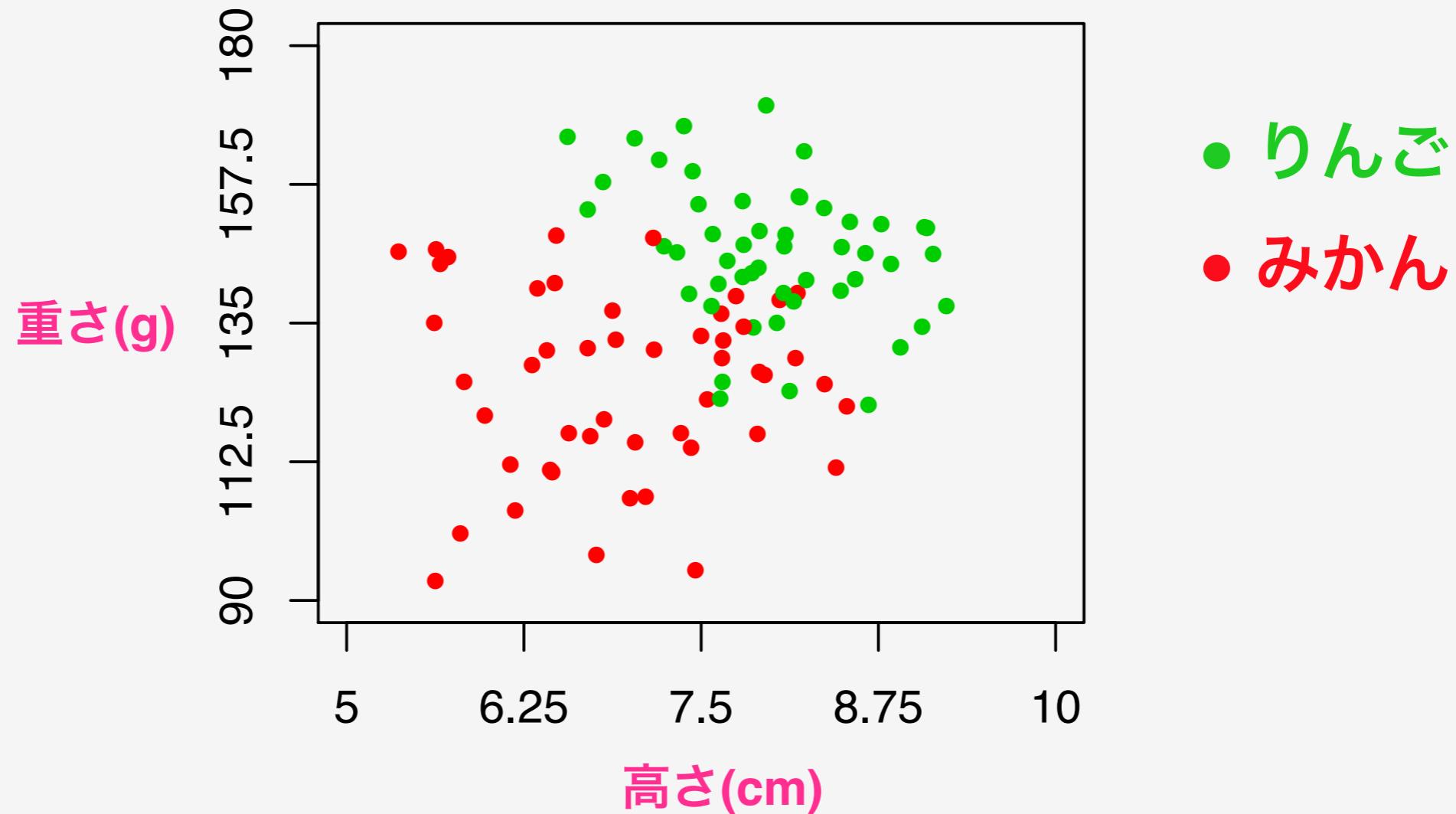
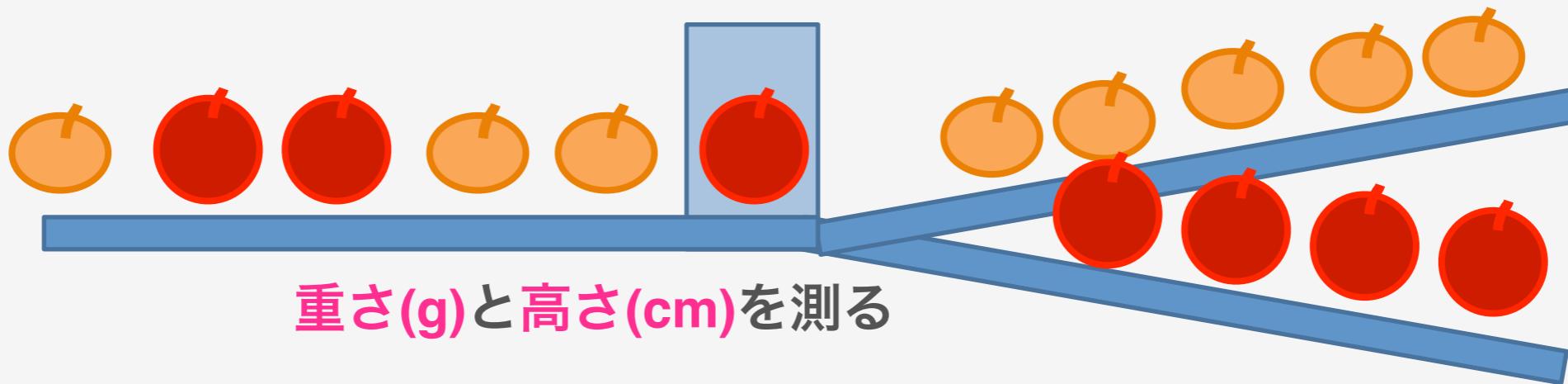
統計的に浮かび上がる法則性



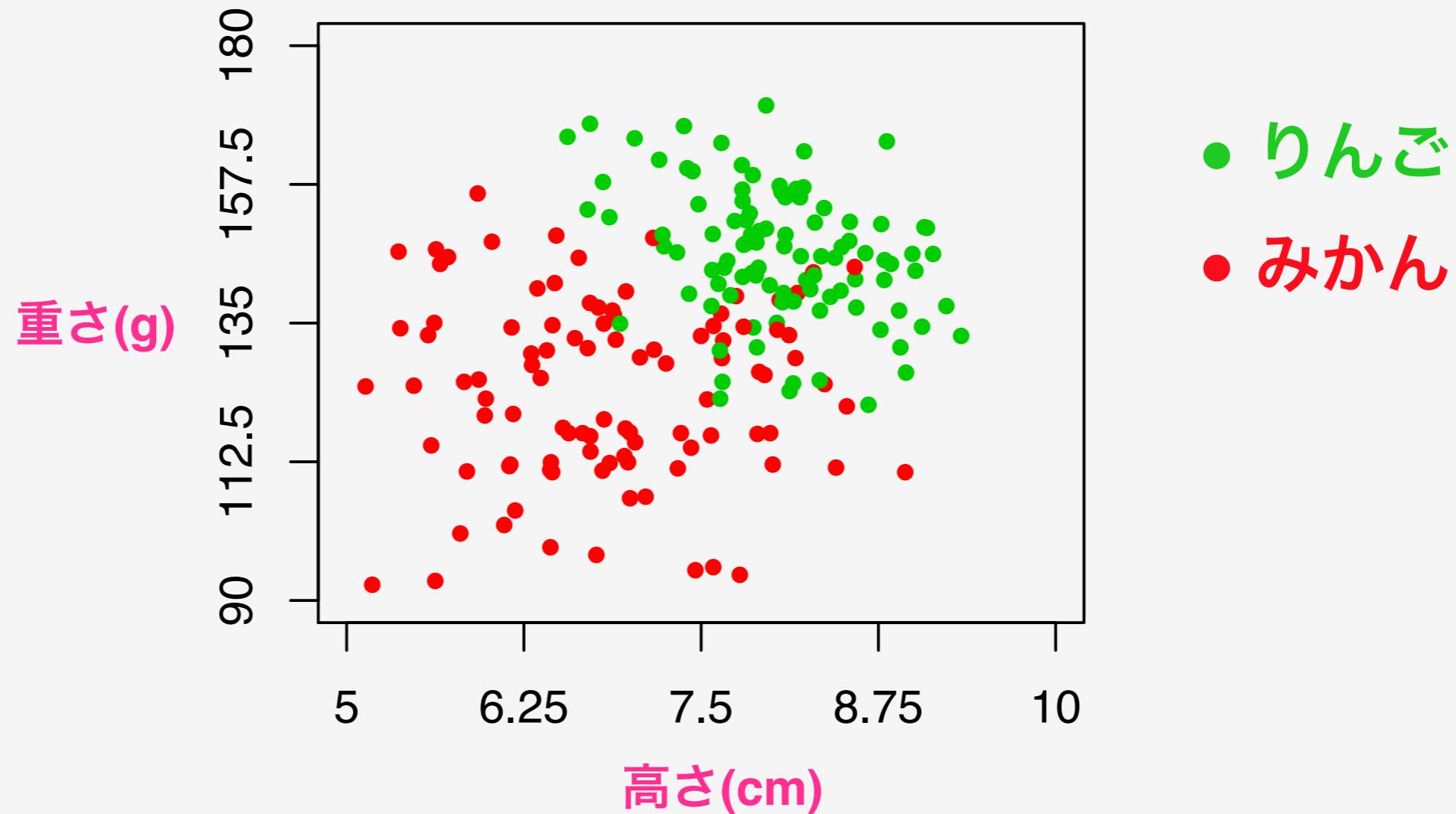
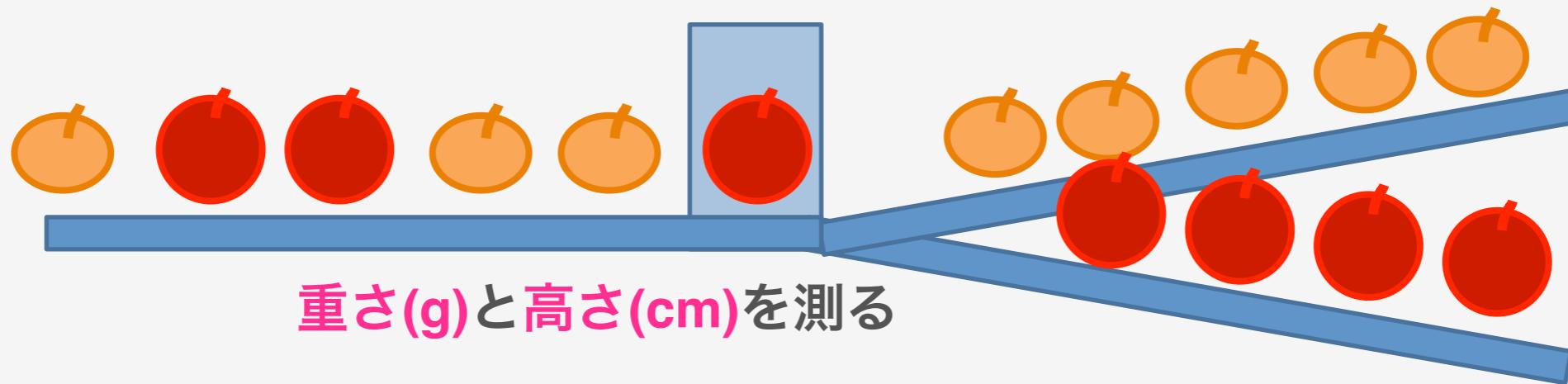
統計的に浮かび上がる法則性



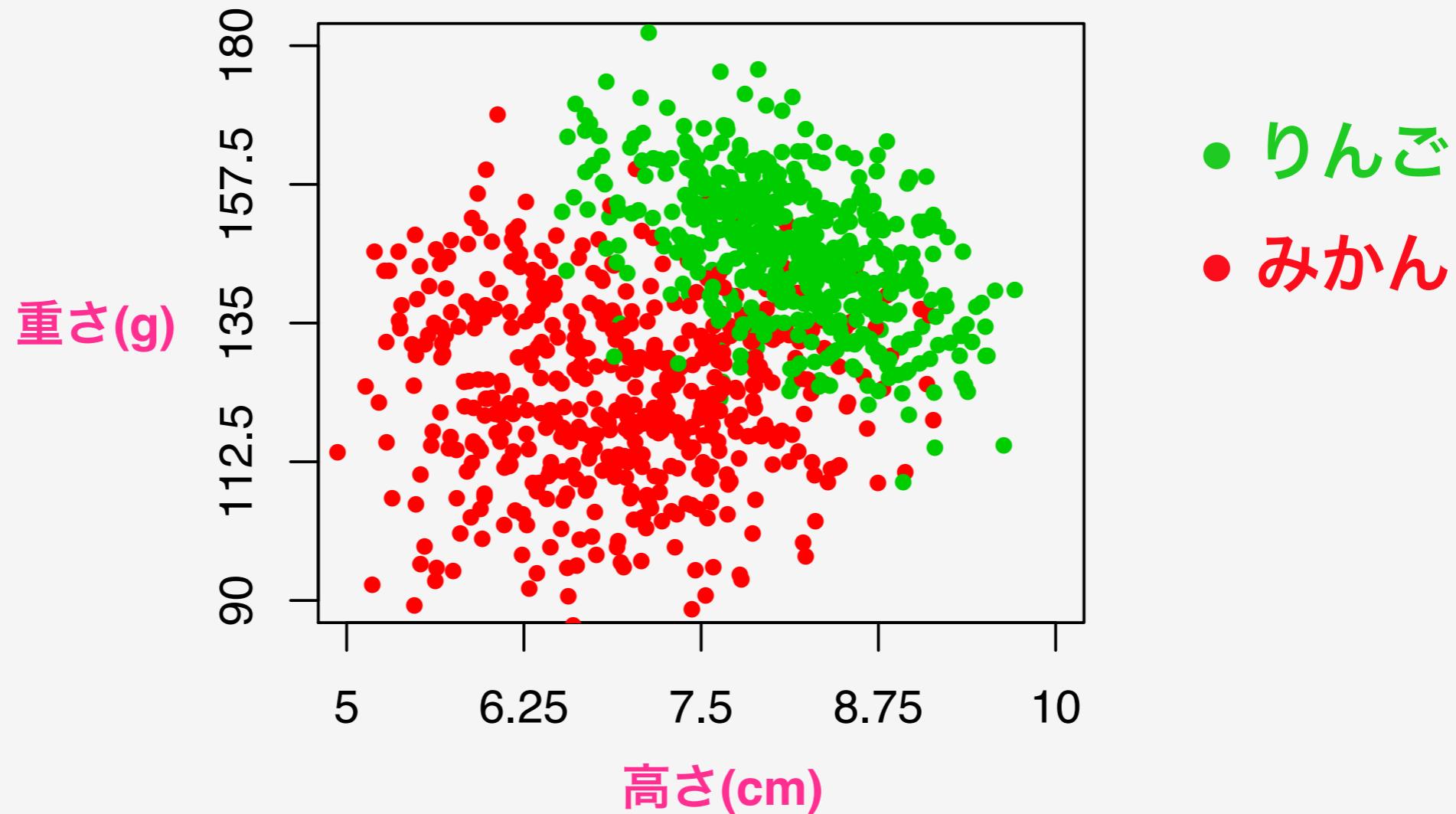
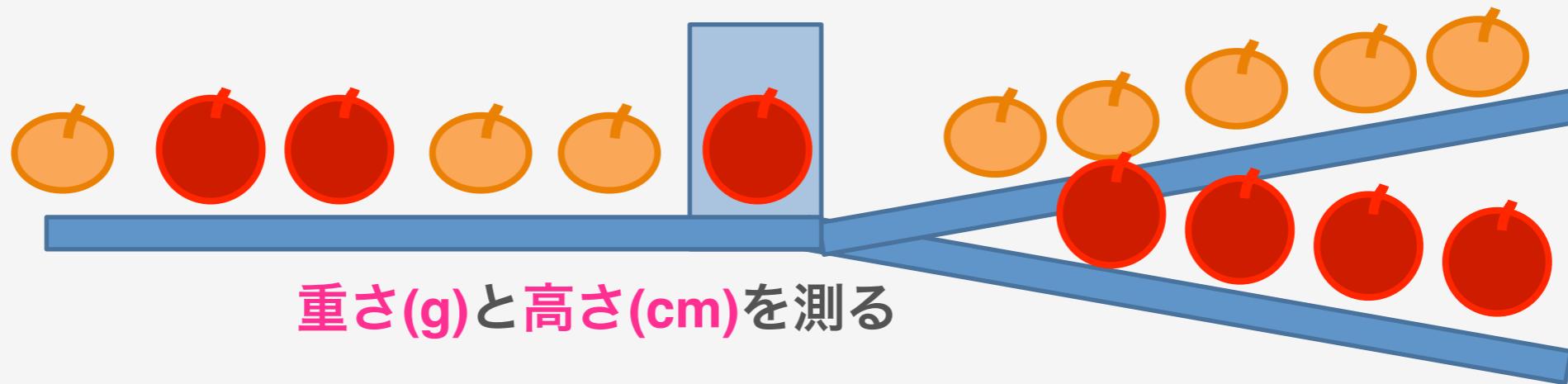
統計的に浮かび上がる法則性



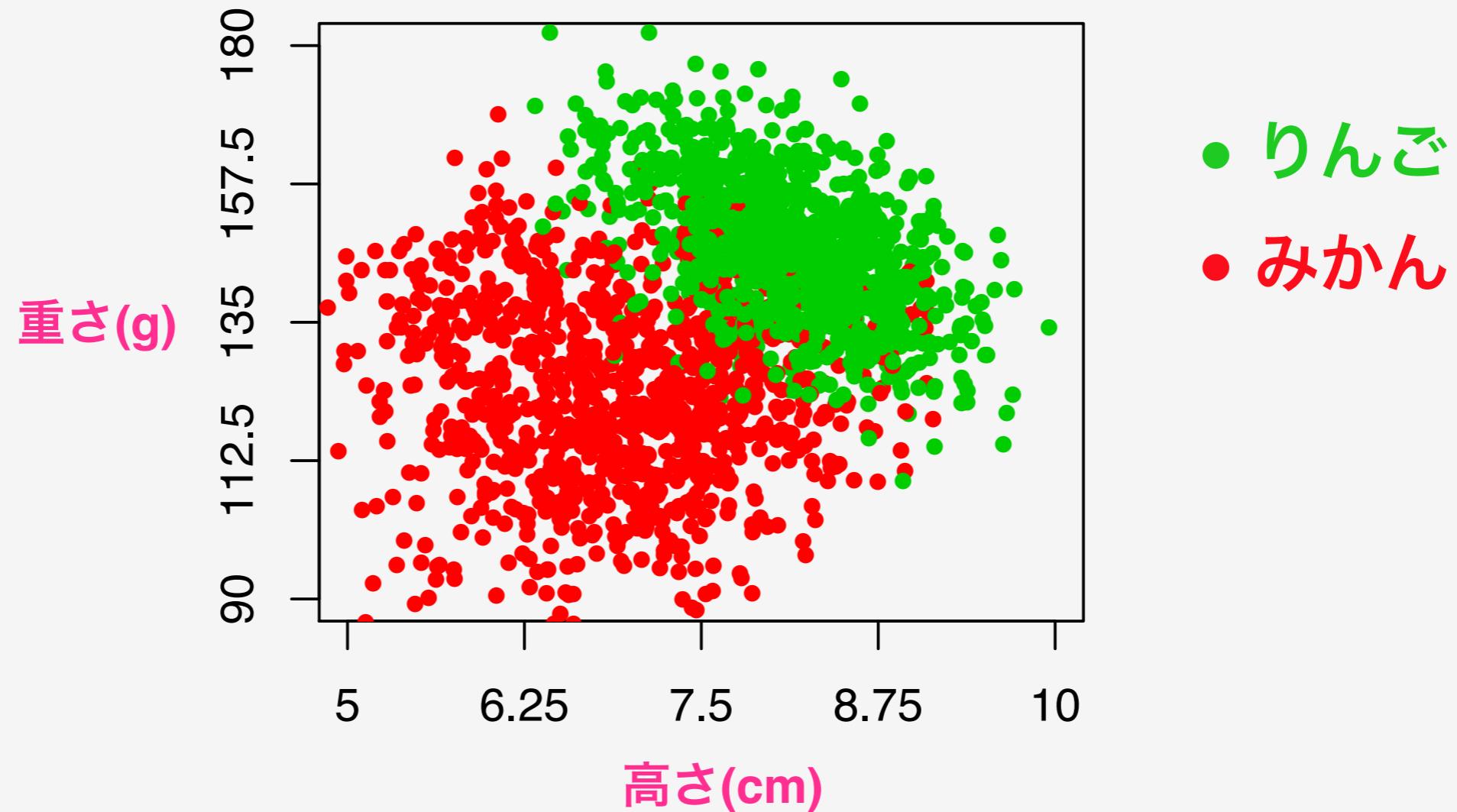
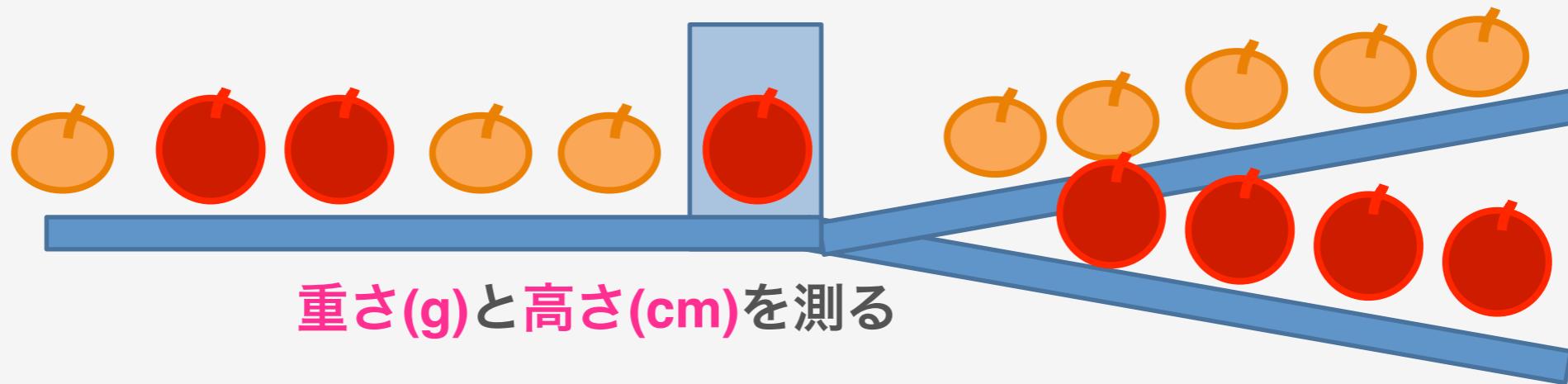
統計的に浮かび上がる法則性



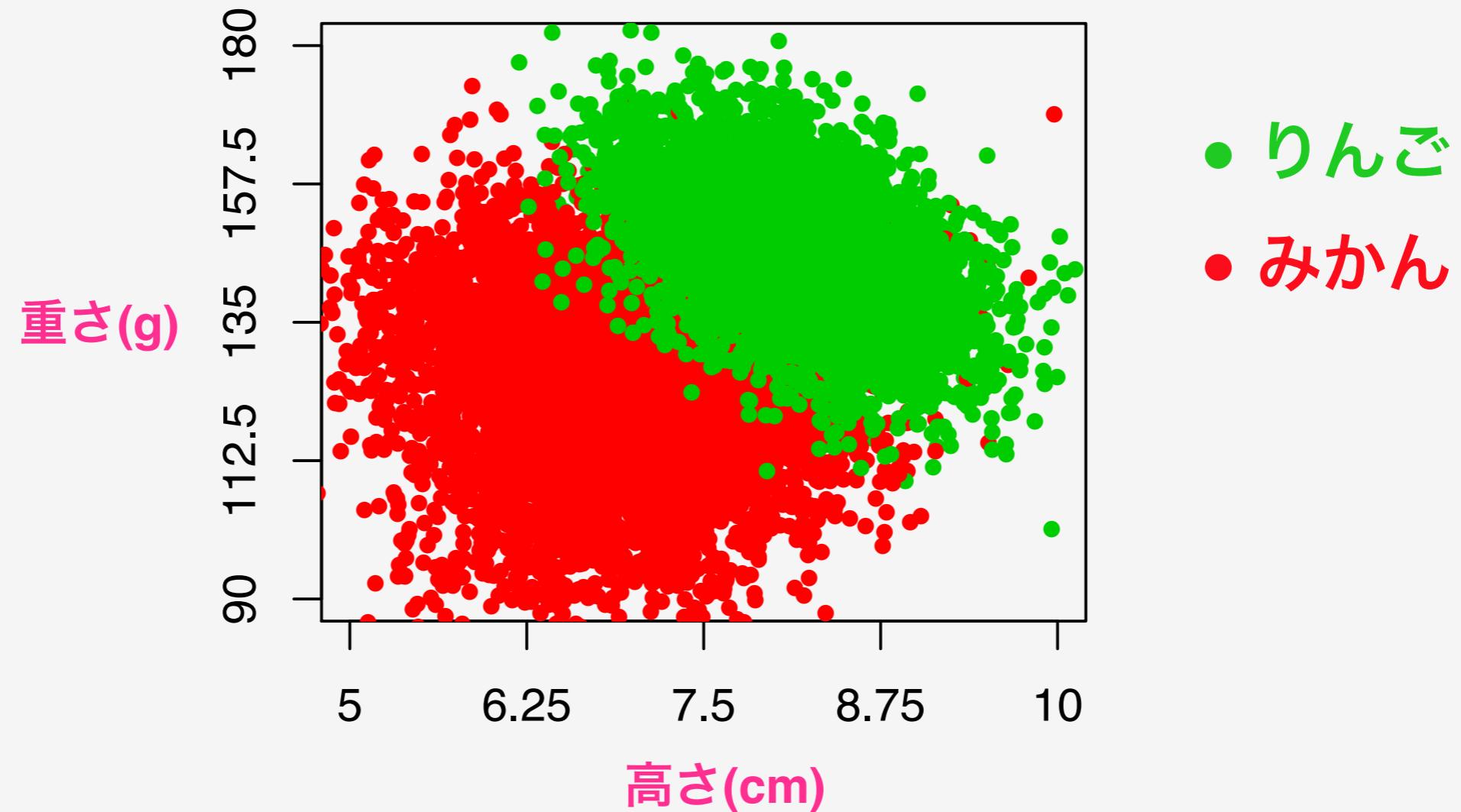
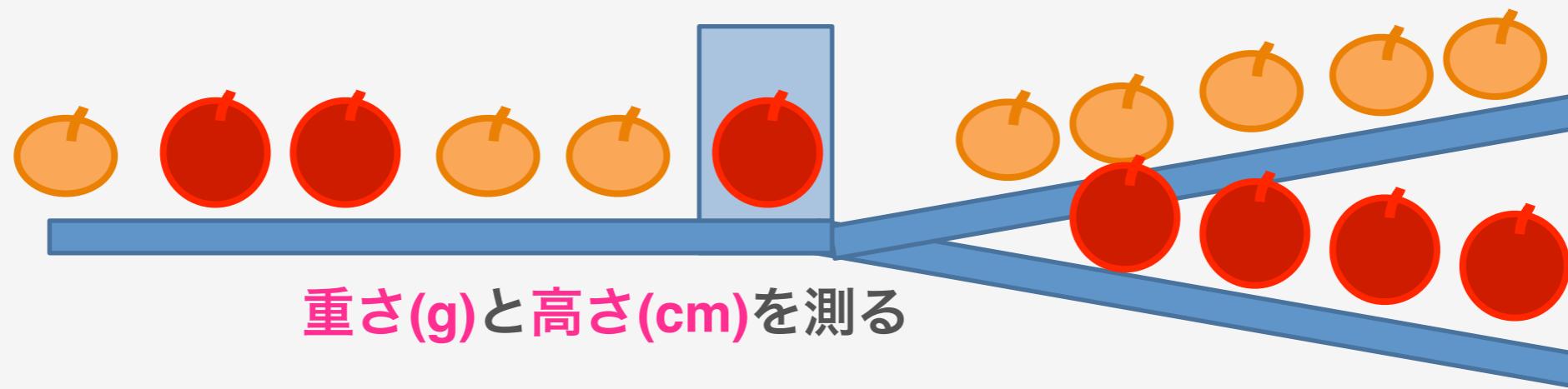
統計的に浮かび上がる法則性



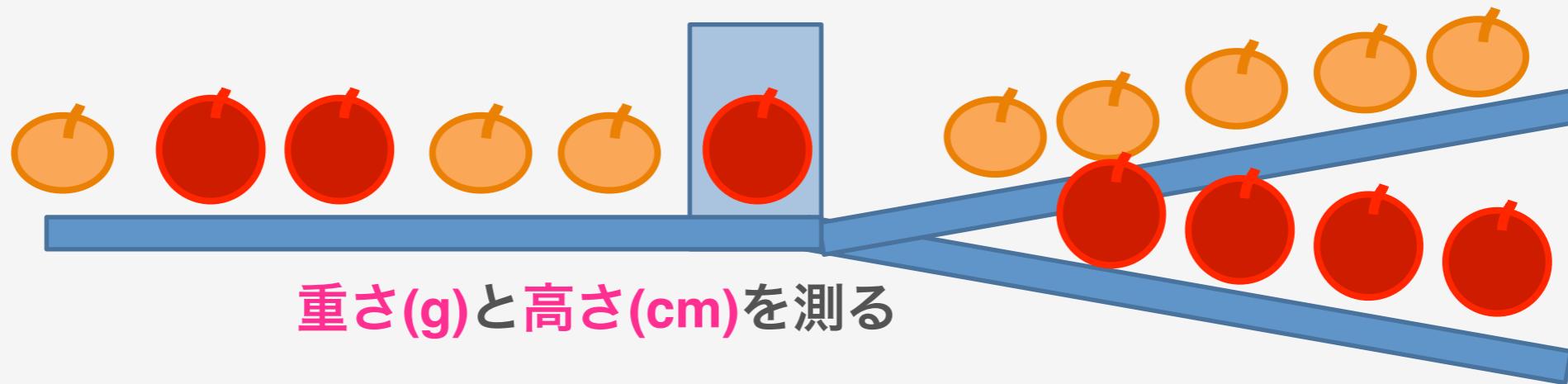
統計的に浮かび上がる法則性



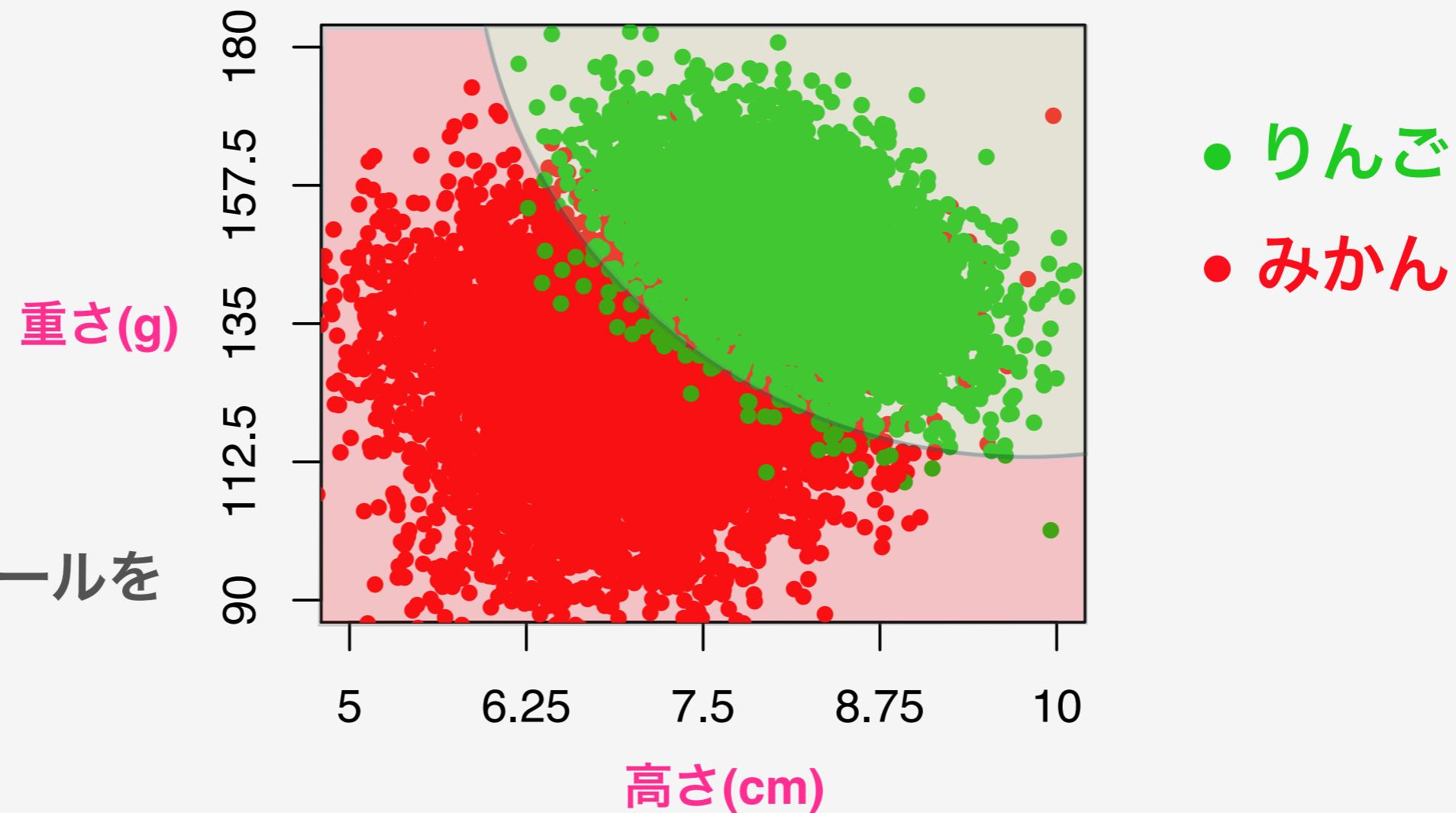
統計的に浮かび上がる法則性



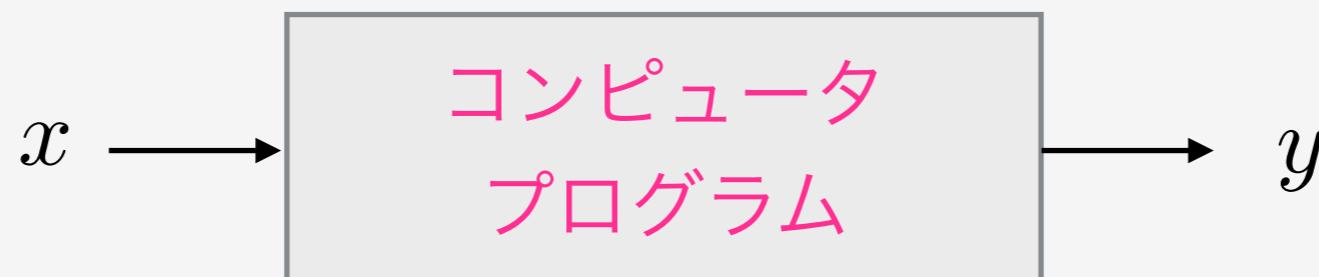
統計的に浮かび上がる法則性



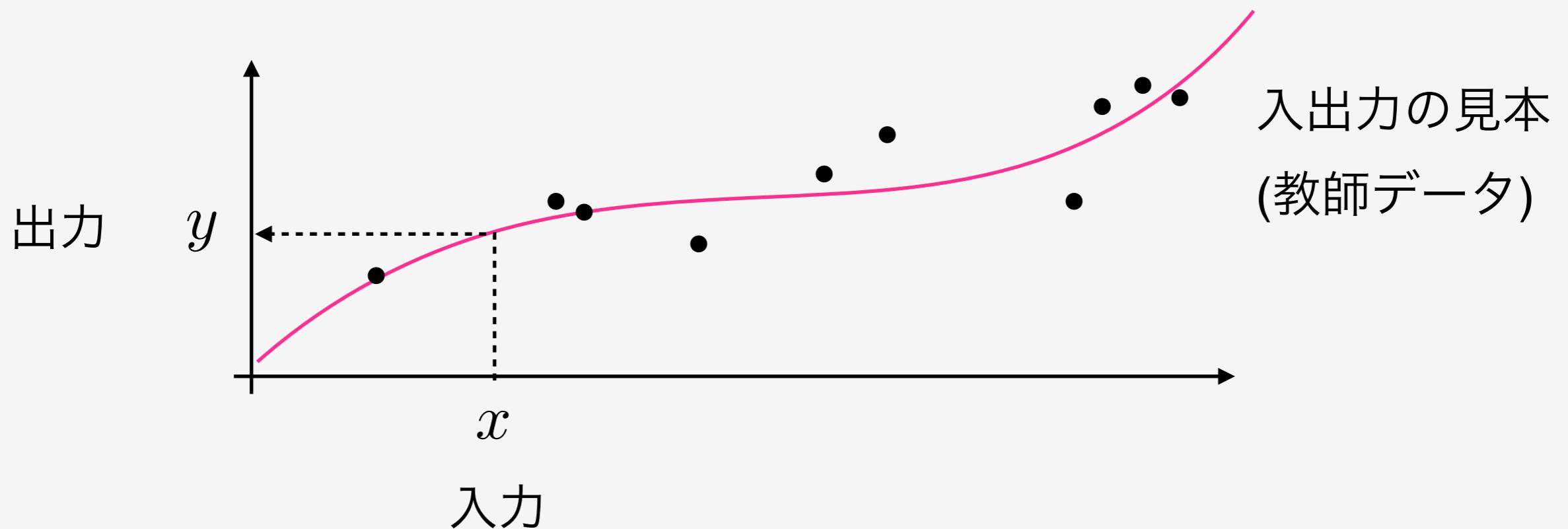
空間を2色で
塗り分けるルールを
決める問題



入力が1次元なら、単に曲線当てはめ問題



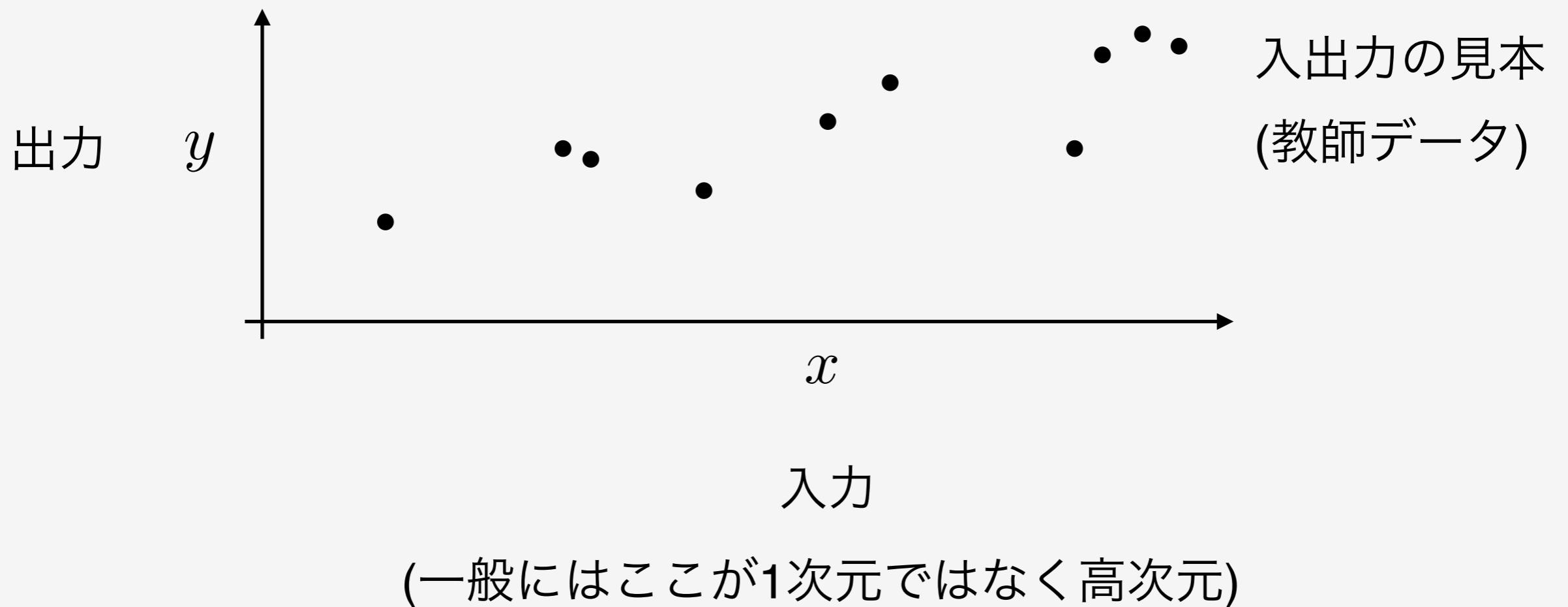
入出力の見本: $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$



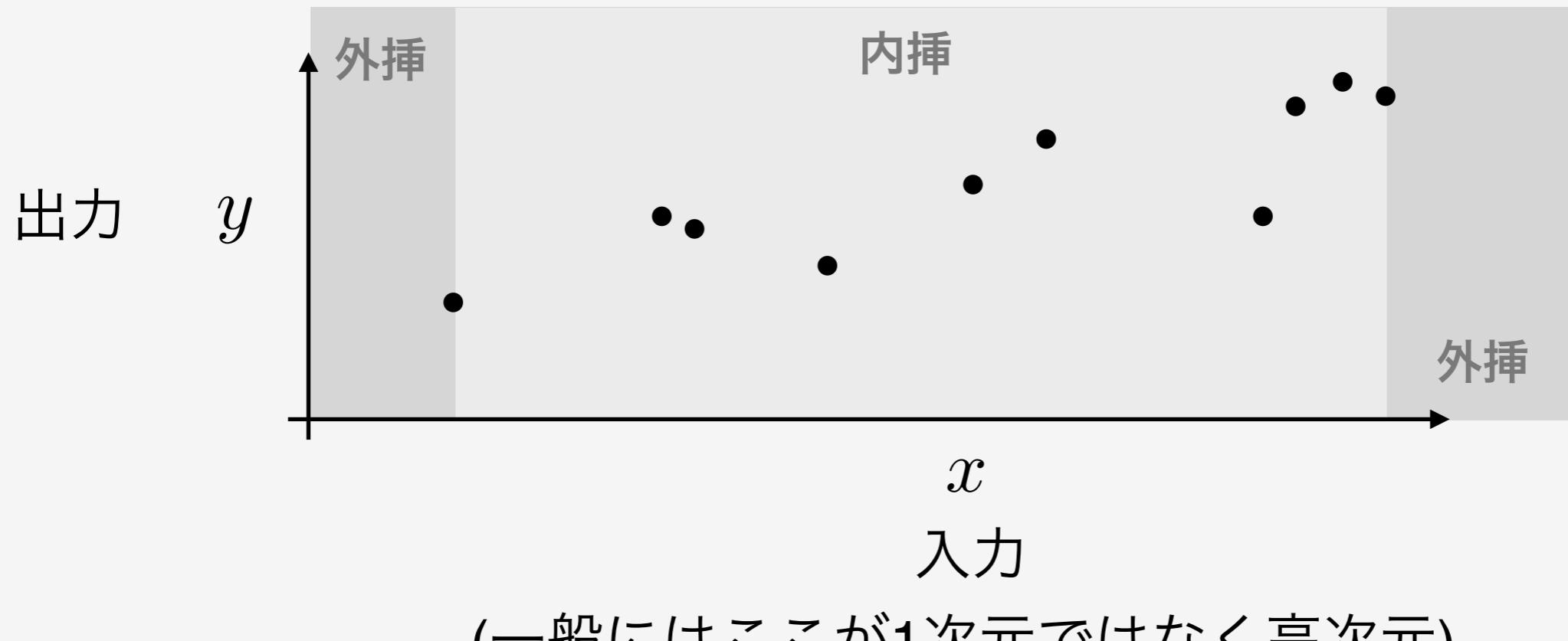
大事な点：現状の教師つき学習の仕組み

手元にあるデータから「手元にまだないデータ」について語るためにやっていることは(非常に高次元での)高度なデータ内挿にすぎない。

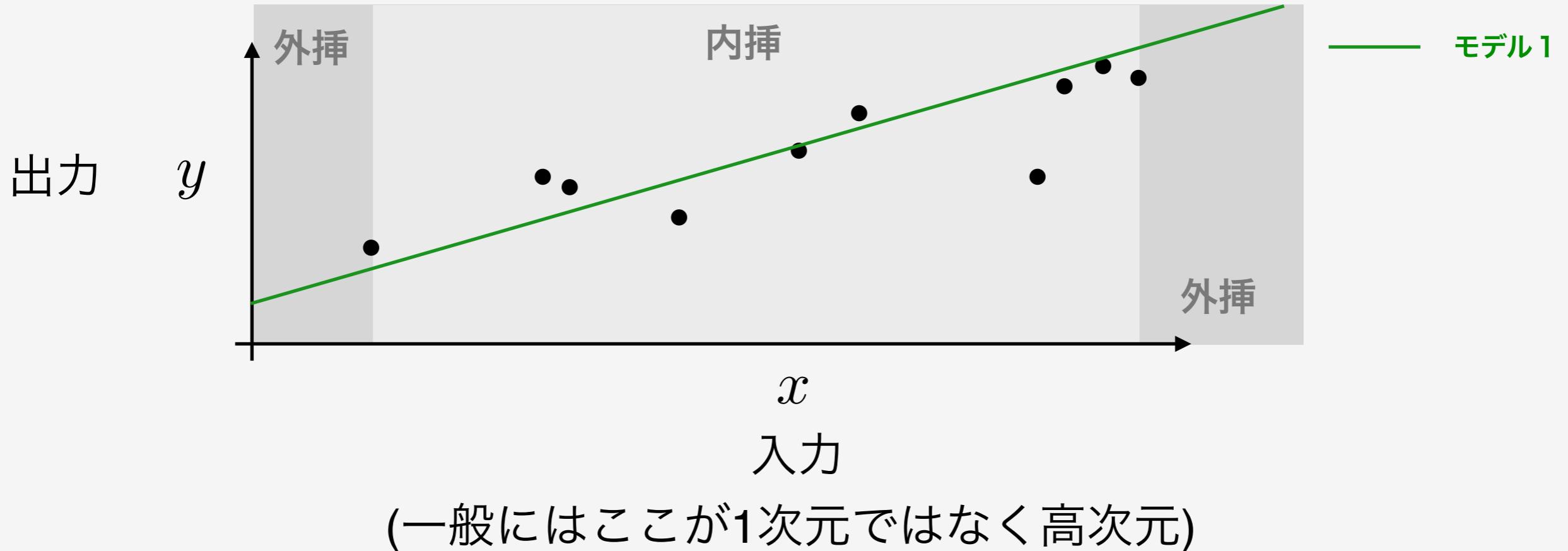
外挿も計算はできるが基本的には不得意(当たらない)



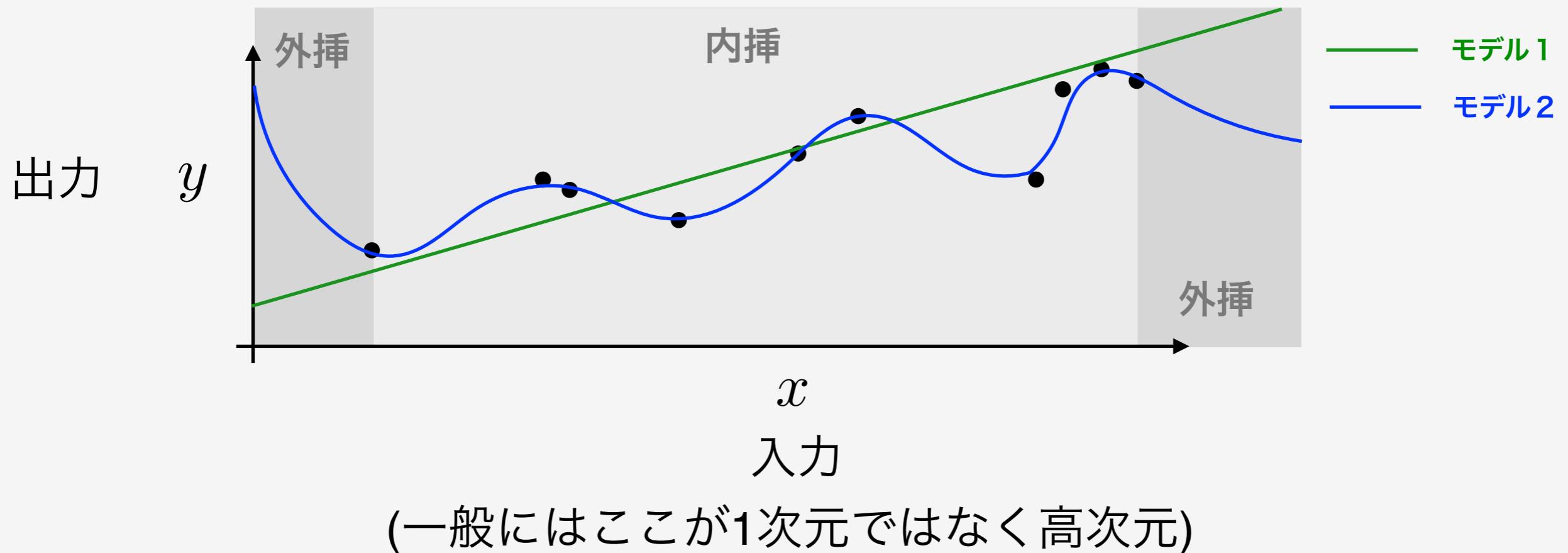
教師つき学習には様々なモデルがある



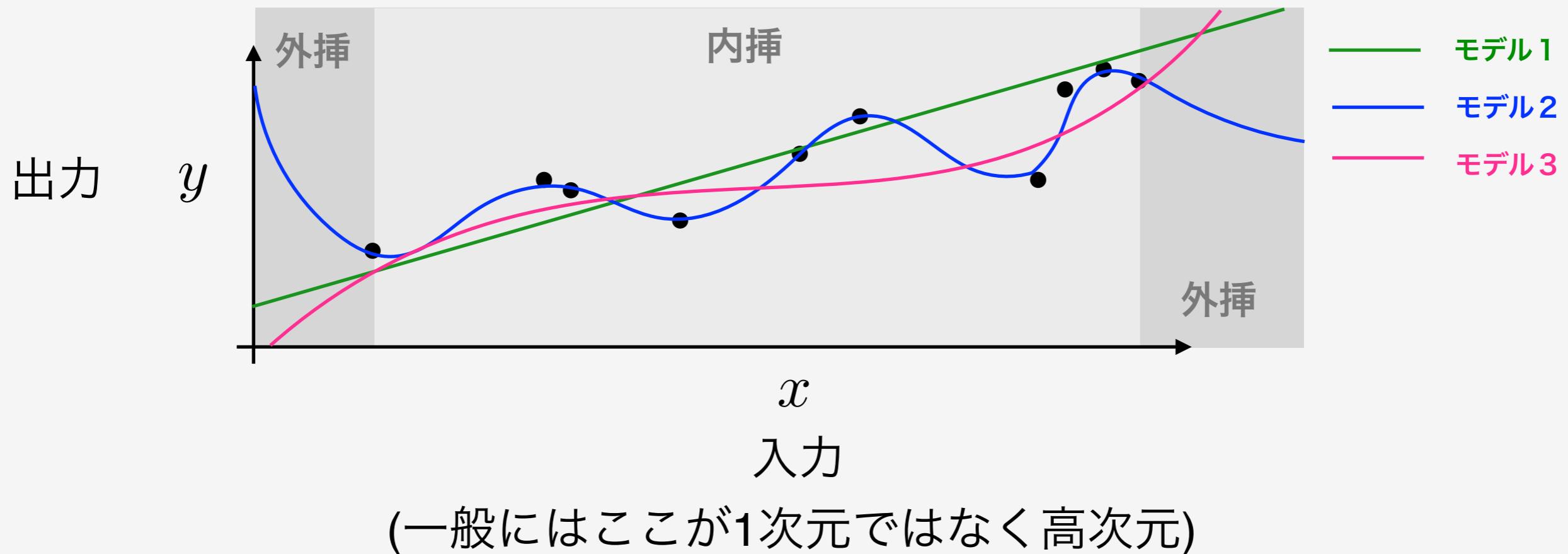
教師つき学習には様々なモデルがある



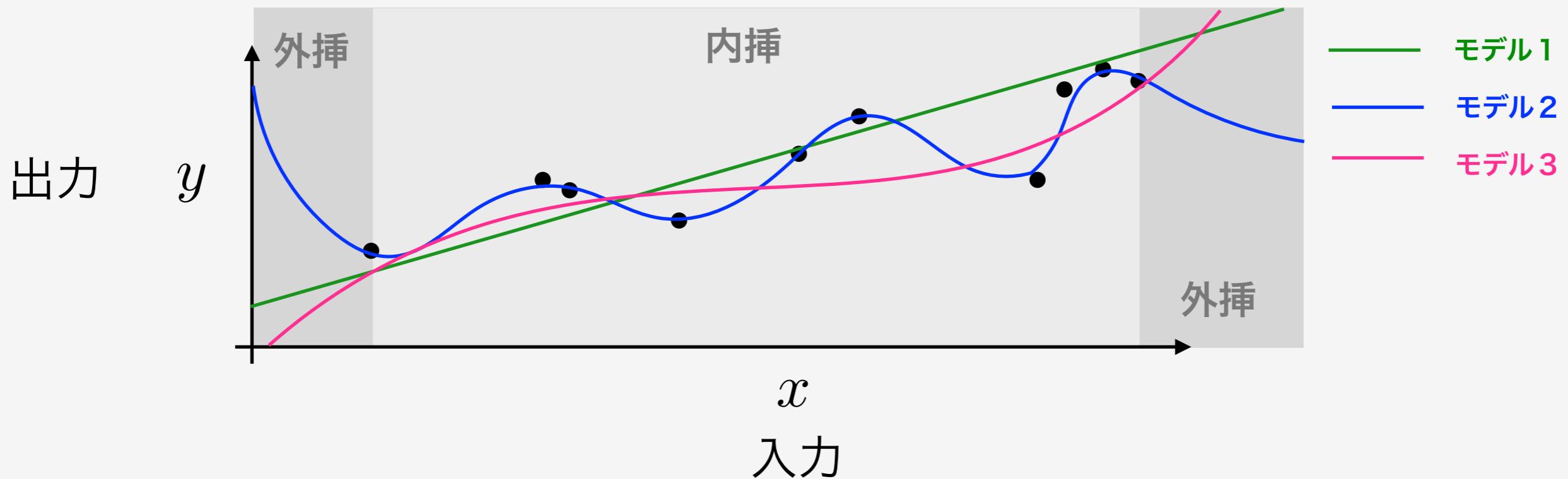
教師つき学習には様々なモデルがある



教師つき学習には様々なモデルがある



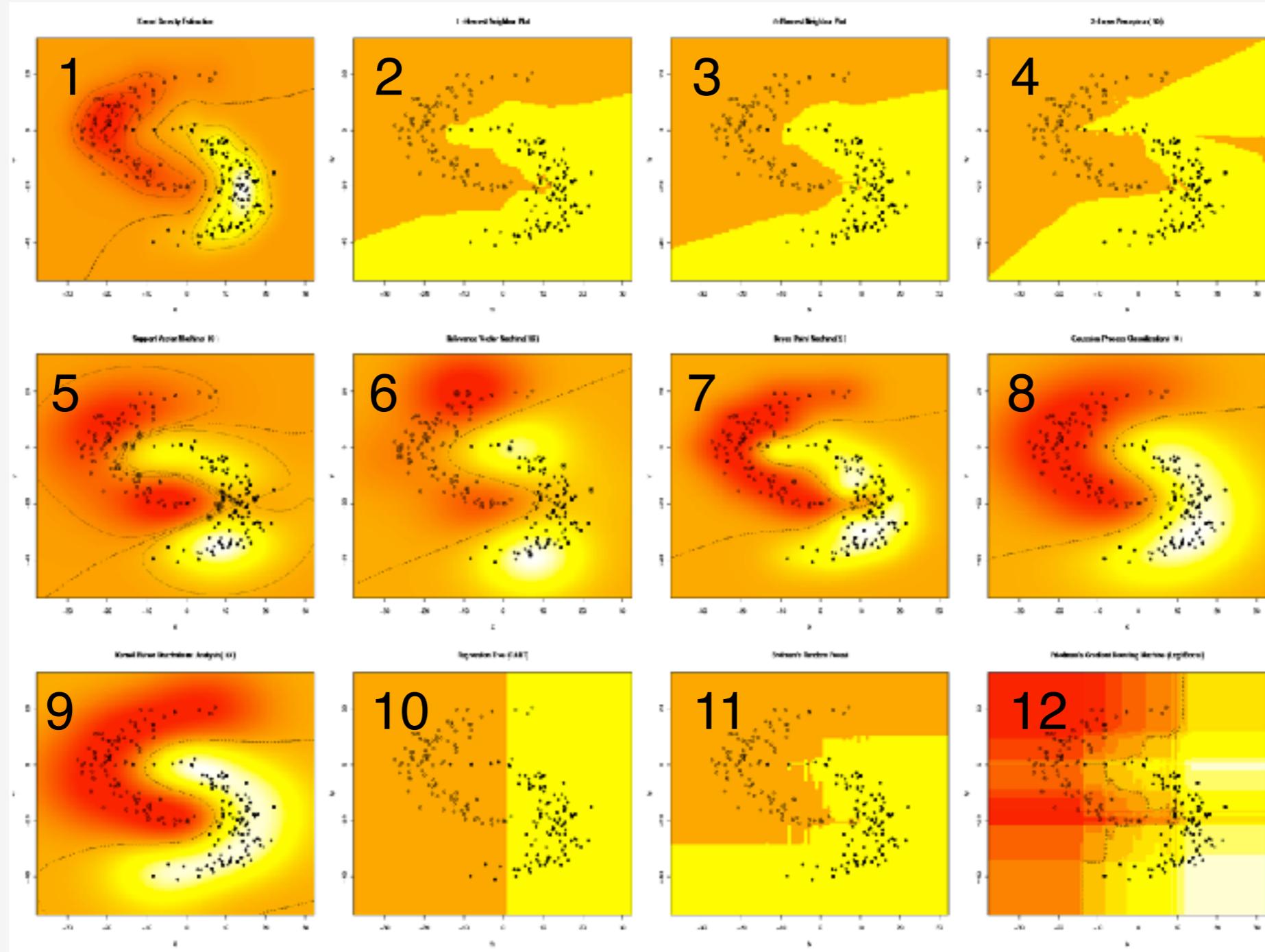
教師つき学習には様々なモデルがある



(一般にはここが1次元ではなく高次元)

- ・ どのモデルや入力表現が良いかは結局問題依存！
- ・ 対象ごとに様々な定式化やノウハウや技術がある
- ・ 出力を説明するのに十分な入力を考へる必要がある。
- ・ 基本的には入力は「決まった長さの」配列（一見そうでなくても）

教師つき学習には様々なモデルがある

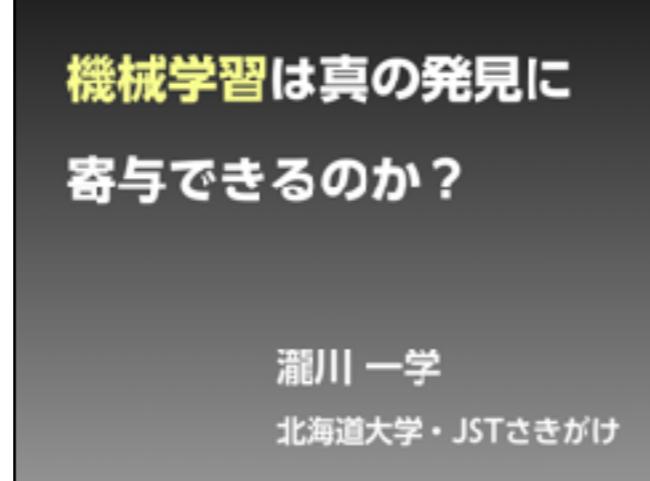


1. Plugin Bayes Classifier
2. 1-Nearest Neighboor Method
3. 5-Nearest Neighboor Method
4. 3-Layer Perceptron
5. Support Vector Machine
6. Relevance Vector Machine
7. Bayes Point Machine
8. Gaussian Process Classifier
9. Kernel Discreminant Analysis
10. Regression Tree (CART)
11. Random Forest
12. Gradient Boosting Machine

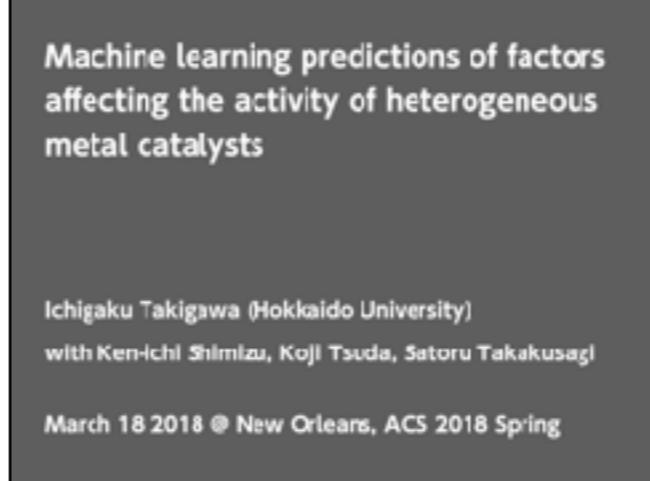
さらに知りたい人 http://art.ist.hokudai.ac.jp/~takigawa/index_ja.html



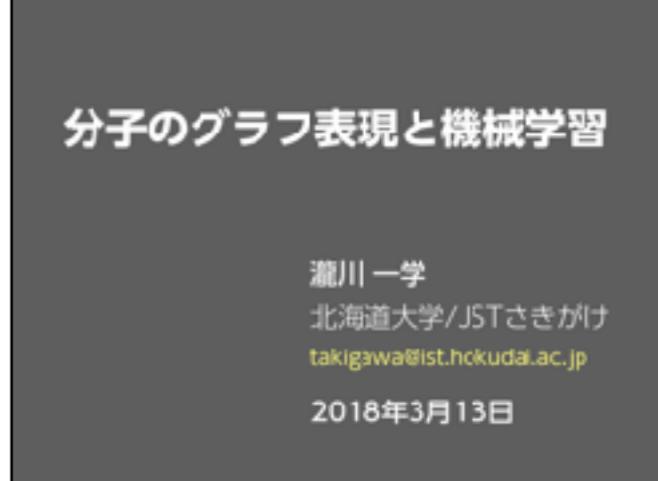
2018.06.07



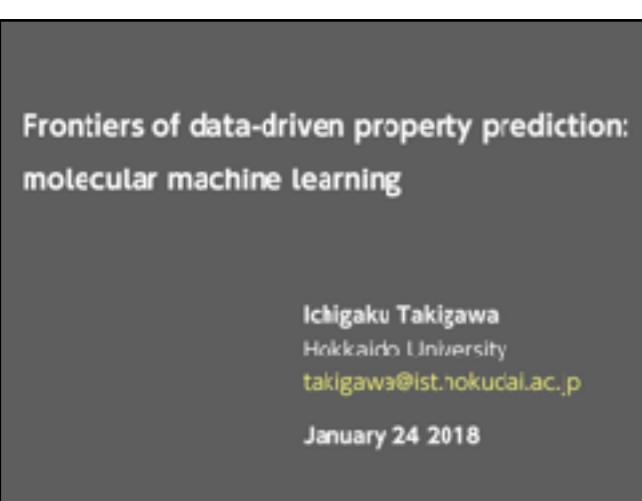
2018.05.21



2018.03.18



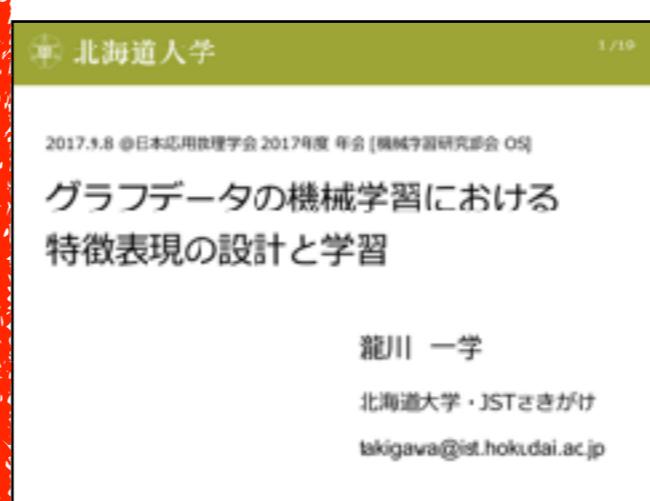
2017.03.13



2018.01.24



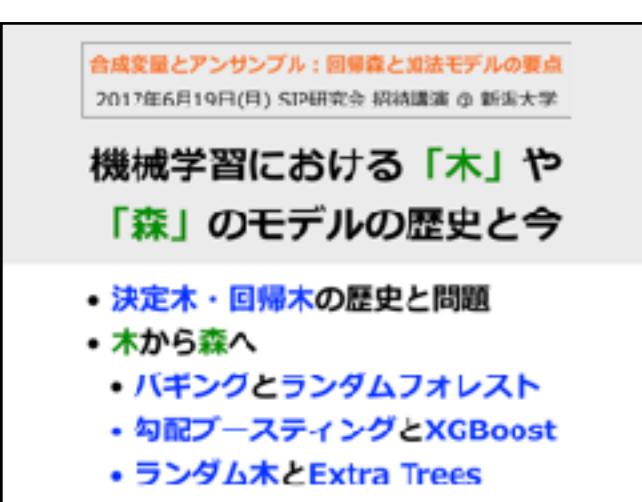
2017.09.08



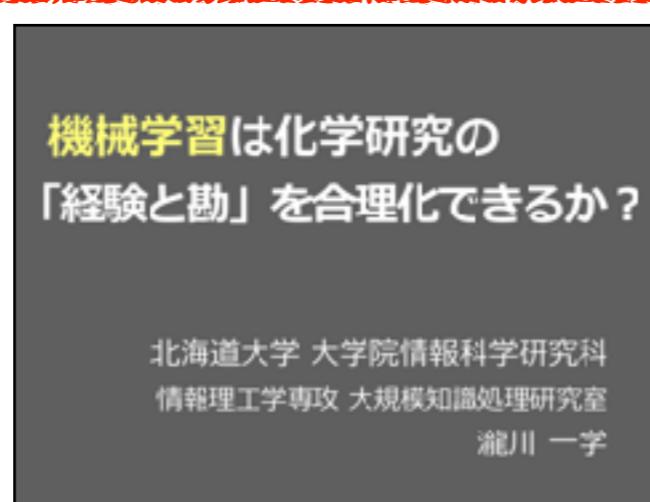
2017.10.01



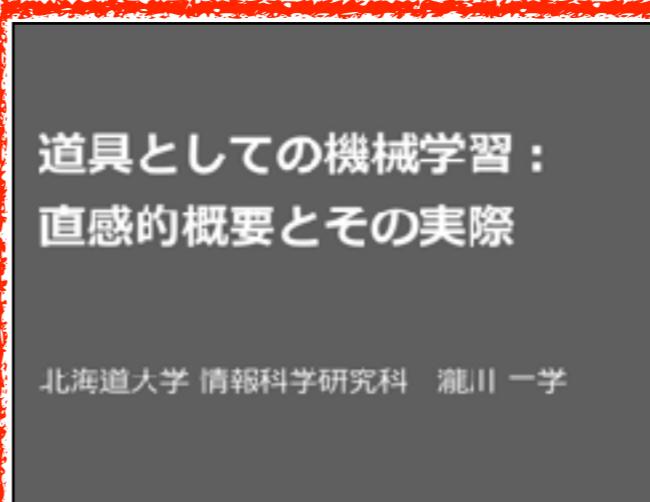
2017.08.01



2017.06.19



2017.06.11



2016.11.19

**作業：授業のホームページから
「practice08.ipynb」をダウンロードして実行してみよう。**



各自なりにここまでの中身を 整理したJupyter Notebookを作る

- あとでそのNotebookをメール提出してもらう予定
- 学びの記録と思って、分からぬことのメモなども含めて各自作成して、それをみながら作業すると定着が早いと思います！
- 新規作成でJupyter Notebookを作成し自由に記述