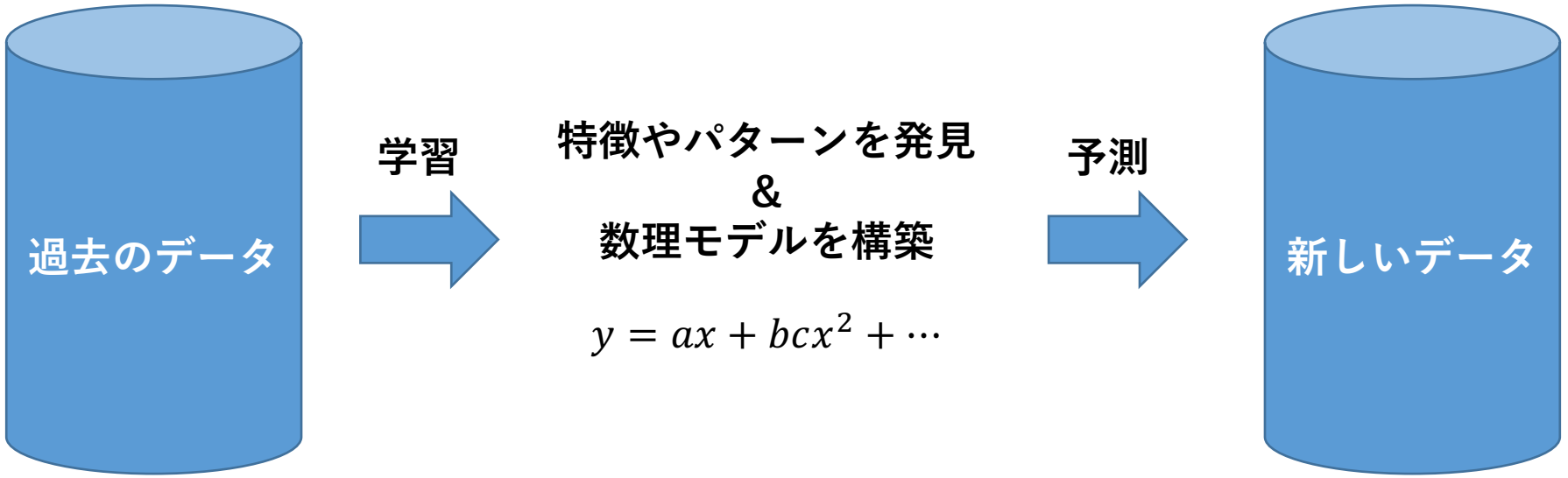
機械学習と統計学の違い
＆
ゲノム解析への応用について

そもそも機械学習とは？

Machine learning (ML) is the study of algorithms and mathematical models
that computer systems use to progressively improve their performance on a specific task...(Wikipedia)



例えば

メヒシバ
Digitaria ciliaris

アキメヒシバ
Digitaria violascens

例えば

＝メヒシバ

＝メヒシバ

＝アキメヒシバ

＝アキメヒシバ

⋮　　⋮

過去のデータから学習

パターンを発見

(ex.アキメヒシバは葉鞘が無毛)

予測

新しいデータ

＝ ？？？

97%の確率でアキメヒシバ
3%の確率でメヒシバ
（多分）

例えば

過去のデータ

ゲノム配列	収穫量
ATTGAC…	2,0kg
GAGGTA…	3,6kg
TGCCGC…	1,1kg
ATCGAA…	2.1kg
…	…

学習

特徴・パターンを発見
数式化
収穫量 $= \boldsymbol{ax} + \boldsymbol{by} + \boldsymbol{cz}$ …

予測

新しいデータ
予測したいデータ

ゲノム配列	収穫量
GAAAAC…	???
TTAGGG…	???
…	…

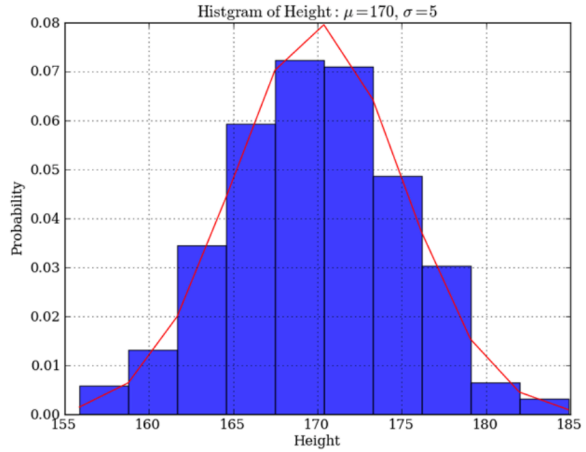
ゲノム配列	予想収穫量
GAAAAC…	1.82kg
TTAGGG…	1.11kg
…	…

統計学と機械学習の違い…使われる手法やアルゴリズムには共通のものも多い

- **統計学**

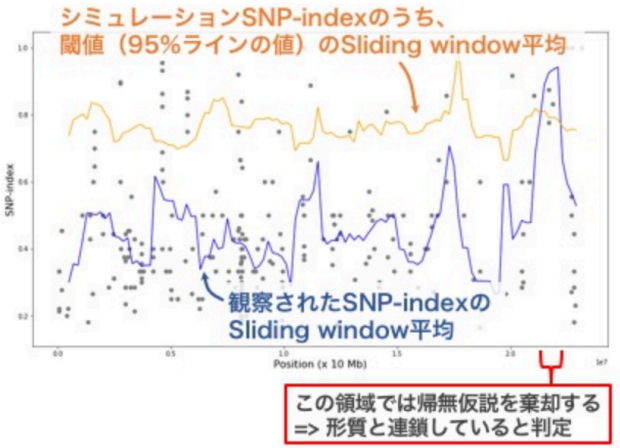
記述統計…統計値や可視化を用いてデータを人間が理解しやすくする

ex) 平均値や分散を計算し、データの性質を知る



推測統計…データに対する考察

ex) 母集団の平均値を推定, SNP-indexで関連している遺伝子の位置を推定



- **機械学習**

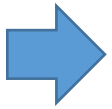
過去のデータから精度の高い予測モデルを作成する

統計学と機械学習の違い…目的・戦略が異なる

- **統計学 …** モデルや手法の選択や妥当性・信憑性が重要

ex) 収穫量 $= ax + by + cz$ …

…精度がそこそこ高い



どの変数が大きな影響を与えているのか？
精度を高くする事が出来た要因は？

ex) aが大きい → xは大きな効果を与える

- **機械学習 …** 如何に高い予測精度を出せるかが重要

→様々な手法、ブラックボックスなモデルであっても予測精度を重視

ex) 収穫量 $= \sqrt[3]{a^8x^{27}z^{55}} + \lim_{n\to\infty}\left(\frac{\int b\sin an}{n}\right)^n + \cdots$

…精度が非常に高い

なぜ精度が高いか理解しにくい場合もある

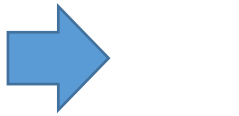
ex) aが大きい → ？？？

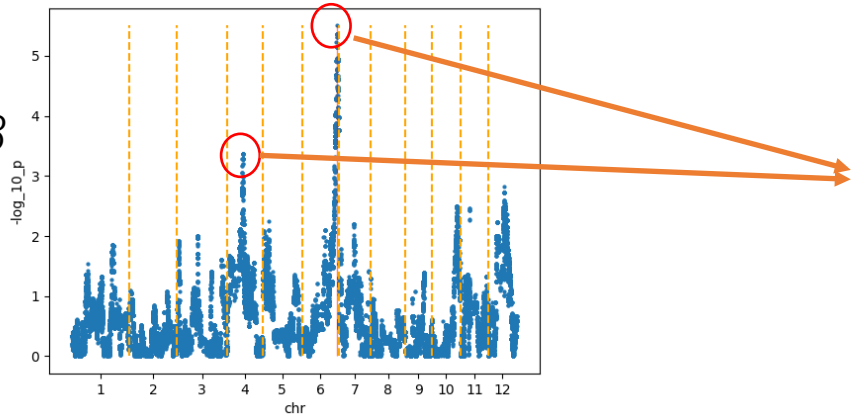
ゲノム解析へ用いる場合

ex1) 遺伝子の機能を知りたい → どれだけモデルの中身が理解しやすいがが重要

収穫量 $= \boldsymbol{\beta_1 gene_1 + \beta_2 gene_2 + \beta_3 gene_3 \ldots}$

各遺伝子の効果β

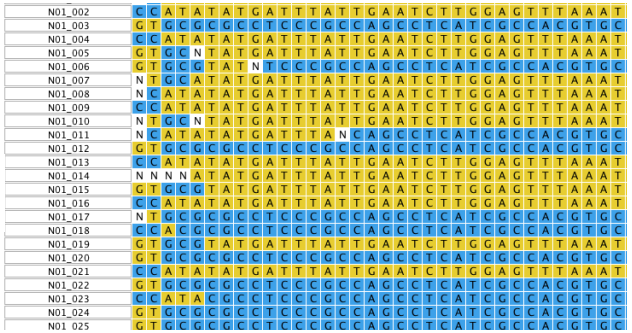




この位置の遺伝子が収穫量に影響を与えている
とモデル式から解釈できる

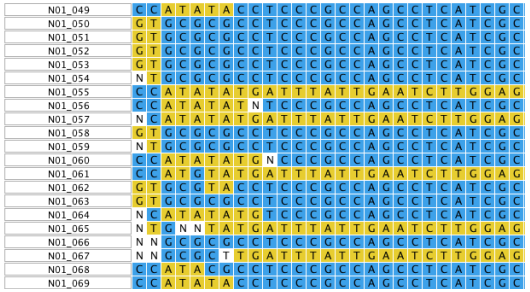
ex2) ゲノム育種などに活用したい → 新しいデータセットでどれだけ正確に予測が行えるか**も**重要

収穫量 ＋



予測モデル

新しいデータ



予測精度

まとめ

- 統計学も機械学習もデータを使って問題解決を図っているが、最終的に求めるゴールが異なる。
- 特にゲノム解析の領域では

  ゲノムデータ → 遺伝子の働く仕組み・進化の機構解明...etc

  ・データが何を意味しているのか

  ・データをもとに作成した数理モデル等から何が読み取れるのか