2019 年 5 月 21 日
フィールドインフォマティクス④
講義スライド

イオノームを用いた機械学習 (分類)

アグリバイオインフォマティクス教育研究ユニット
特任准教授
大森 良弘

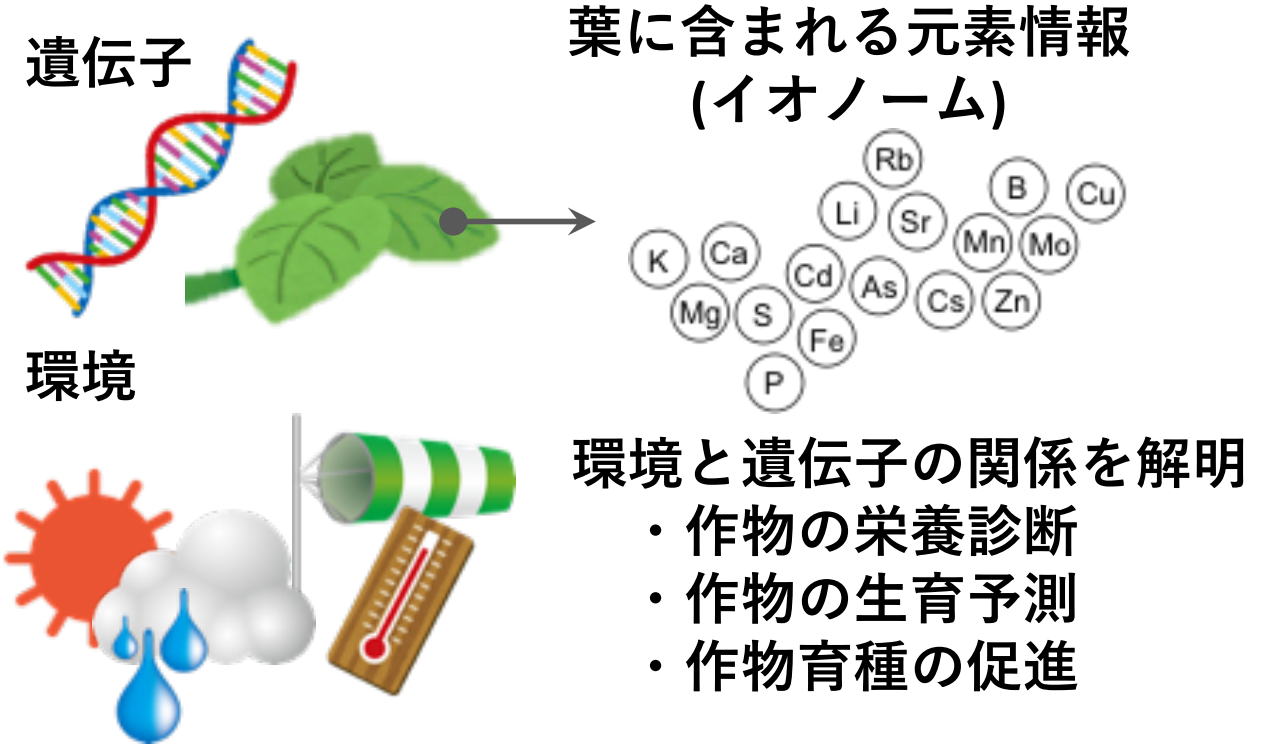
自己紹介

(e-mail: ayohmori@g.ecc.u-tokyo.ac.jp)

名前：大森 良弘 (おおもり よしひろ)

職業：アグリバイオインフォマティクス教育研究ユニット
特任准教授

研究分野：植物分子生物学、植物育種学、植物栄養学、
フィールドインフォマティクス



講義の目的

- 「**こういうコマンドでこういうことができるんだ**」がわかることが目的

本講義では、R や機械学習を用いて「何ができるか」の理解を深めることを優先します。
機械学習のアルゴリズムに関する理論的背景は説明しません。

- 「**探索的データ解析により仮説を立て、予測モデルを構築する**」ことが目的

探索的データ解析なしに優れた機械学習モデルは構築できません。
本講義では「機械学習 はじめの一歩」として、 データの視覚化や相関性の解析など、予測モデルの構築に必要な探索的データ解析への理解を深めます。

講義の流れ

R でデータの把握・可視化

- 使用するデータ
- データの型
- 基本統計量
- 箱ひげ図
- ヒストグラム
- 相関

R で機械学習

- 決定木
- ランダムフォレスト
- 多次元尺度構成法
- 部分従属プロット
- 交差検定
- R Markdown でレポート作成

講義で使うデータの紹介 (栽培条件・品種)

鳥取大学乾燥地研究センターの砂質圃場で栽培
(2018 年 6 月中旬から 8 月下旬) した大豆の実際の元素データ

C または D の条件で 4 種類 (品種) の大豆を栽培

タチユタカ、赤莢、白三ツ豆、ホージャク喰不

大豆品種の情報 (農業生物資源ジーンバンク)
https://www.gene.affrc.go.jp/databases-core_collections_jg.php

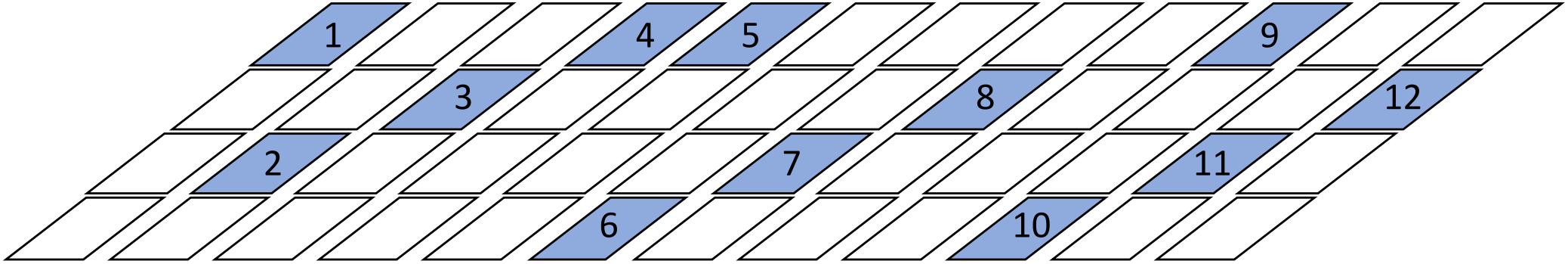
講義で使うデータの紹介 (栽培個体数)

1 区画 (plot) に 4 個体栽培

コントロール環境
１品種１２区画 = ４８個体
４品種合計 = １９２個体

C と D の合計 = ３８４個体

コントロール環境圃場全体 (48 区画)



各品種はランダムに配置して栽培

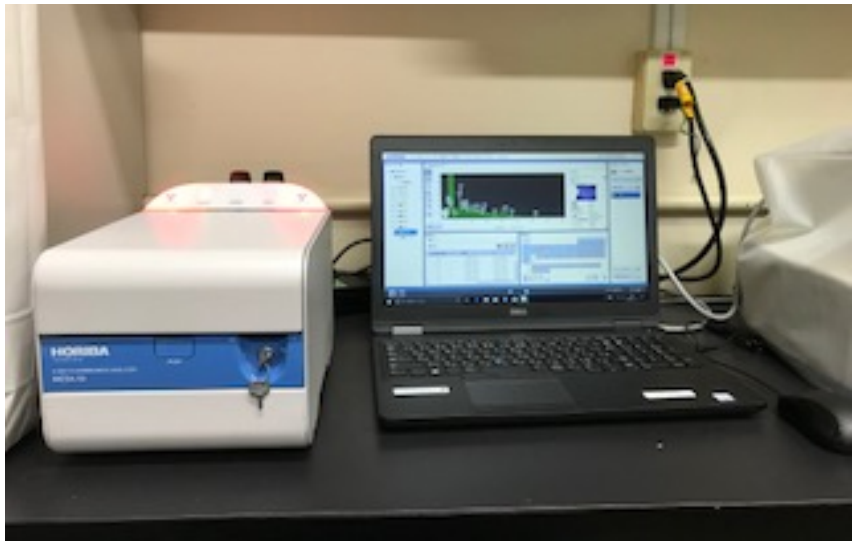
講義で使うデータの紹介 (元素データと新鮮重量)

蛍光 X 線元素分析装置

MESA-50 HORIBA Scientific

蛍光X線元素分析装置

SDD検出器搭載で迅速分析。
一人で持ち運び可能・
バッテリー駆動で
どこでも手軽に分析できます。



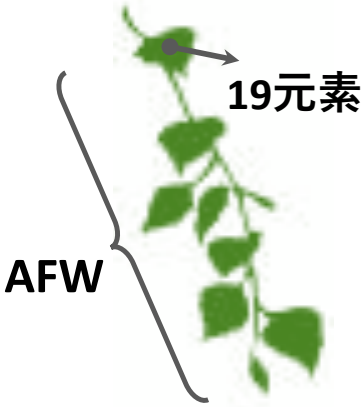
測定可能な元素 (ICP-MS との比較)

ICP-MS

Li, B, Na, Mg, P, S, K, Ca, Mn, Fe, Co, Ni, Cu, Zn, As, Rb, Sr, Mo, Cd, Cs

蛍光 X 線分析装置 (XRF)

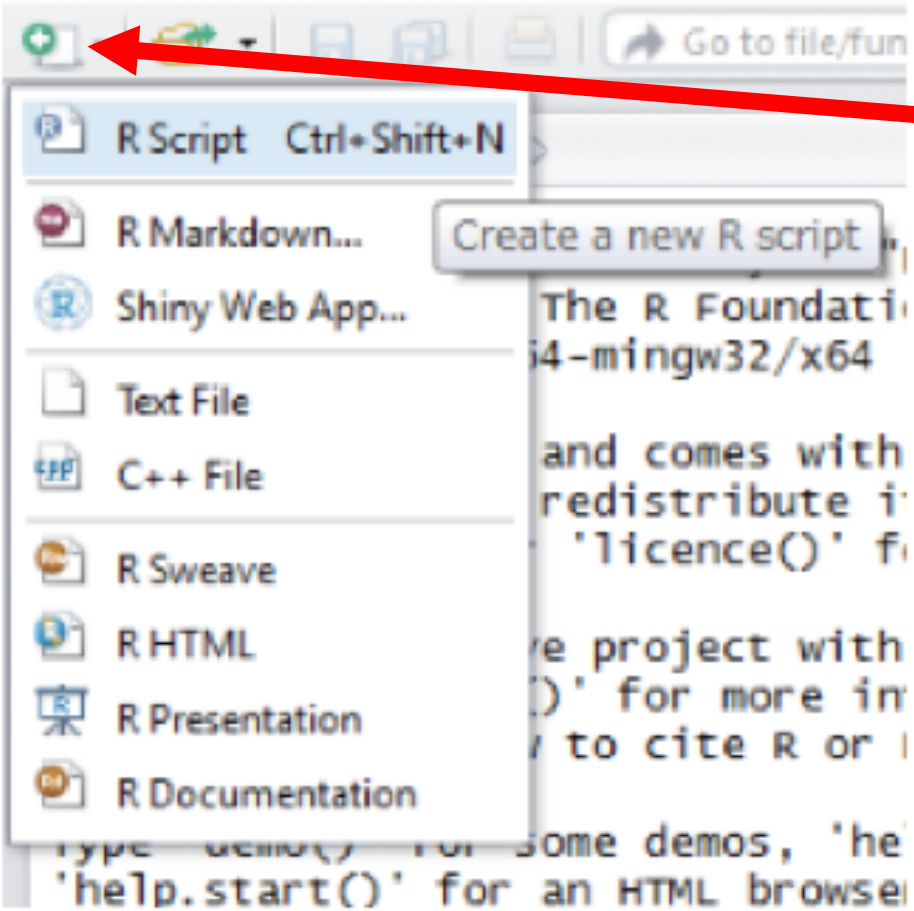
Si, P, S, Cl, K, Ca, Mn, Fe, Co, Ni, Cu, Zn, As, Br, Rb, Sr, Mo, Cd, Cs



8 月下旬に測定された大豆の葉の 19 元素と大豆の新鮮重量 (AFW: 地上部バイオマス) のデータを使用

1 実演

RStudio を開いてデータを読み込んでみよう



Rstudio を開いてメニューの
左上にある
新規ファイル作成ボタンから
「R Script」をクリック

1 実演

RStudio を開いてデータを読み込んでみよう

スクリプトやファイルの編集
R オブジェクトの表示

オブジェクト一覧
実行したコードの履歴な
どを表示

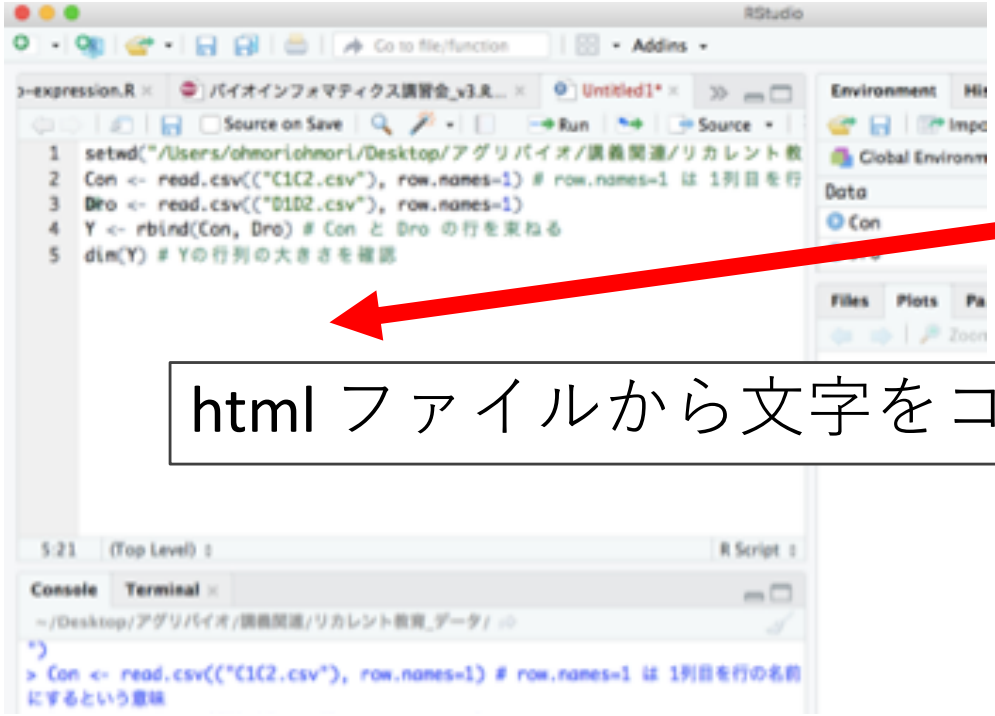
Rコンソール
ここで直接処理を
実行することも可能

ファイル一覧やグラフィクス
ヘルプなどを表示

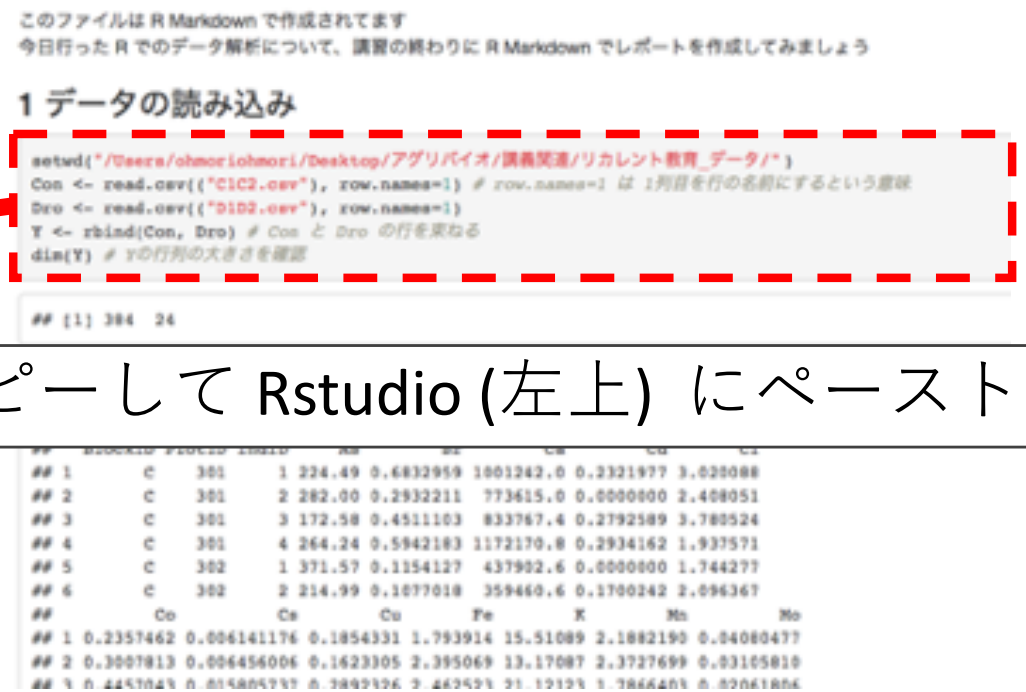
1 実演

RStudio を開いてデータを読み込んでみよう

Rstudio



html ファイル



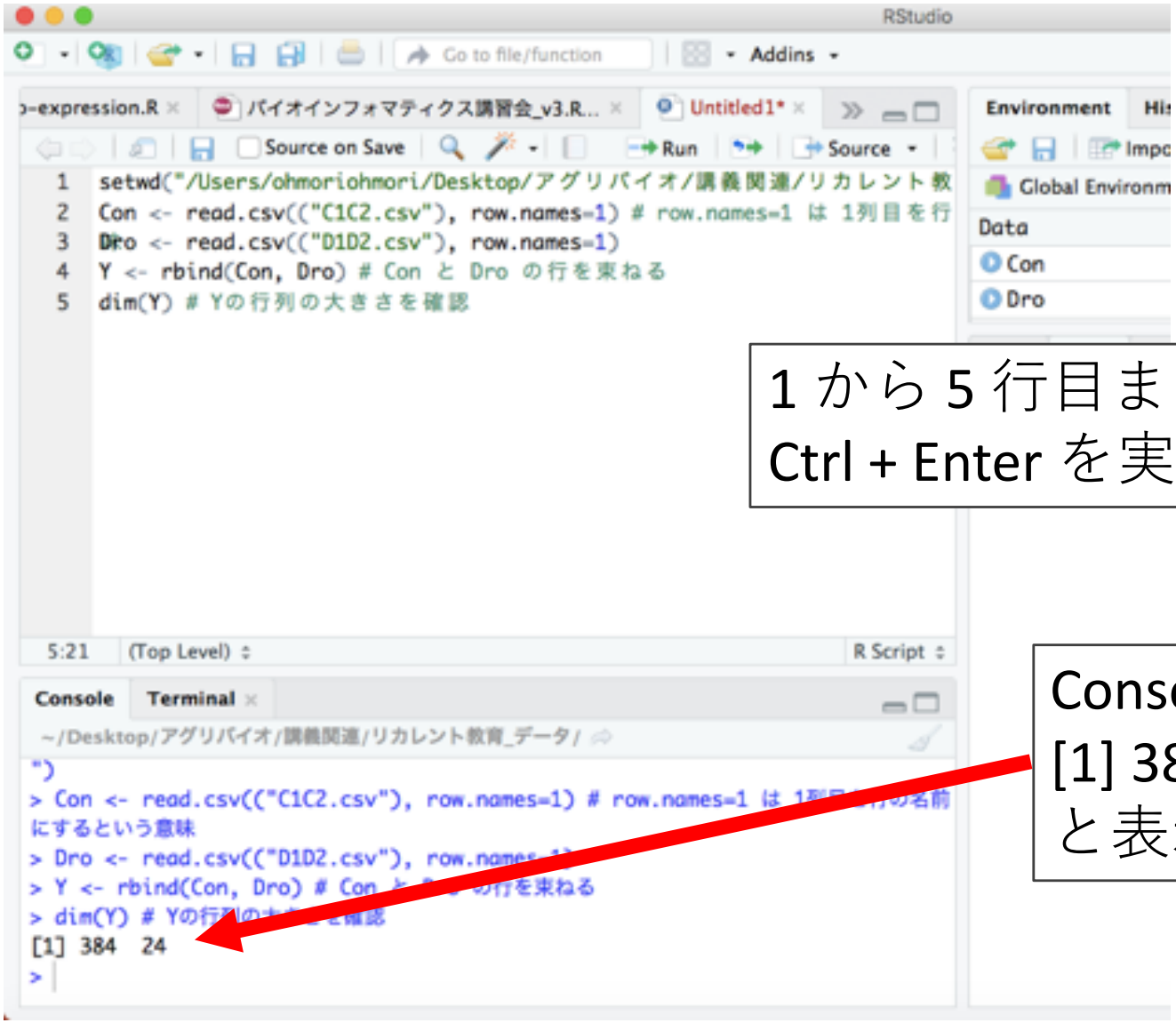
html ファイルから文字をコピーして Rstudio (左上) にペースト

```
setwd("/Users/ohmoriohmori/Desktop/アグリバイオ/講義関連/リカレント教育_データ/")
```

この赤字の部分を、自分の PC のデータが置いてあるフォルダに変更

1 実演

RStudio を開いてデータを読み込んでみよう



1 から 5 行目まで、
Ctrl + Enter を実行

Console に
[1] 384 24
と表示されたら成功

2-1 実演

データの型

- 数値型 (numeric)　例：3, 1.23
- 整数型 (integer)　例：1, -2, 3
- 論理型 (logical)　例：TRUE, FALSE
- **文字型 (character)**　例："apple", ”windows"
- 複素数型 (complex)　例：1.23+0.45i
- **因子型 (Factor)**

因子型は、文字列型を擬似的な数値として扱う質的変数

```
$ BlockID: Factor w/ 2 levels "C","D": 1 1 1 1 1 1 1 1 1 1 ...
```

Levels: “C”, “D” : 111111…と表示されているが、
これは「**C に 1 という数字を当てはめています**」という意味

統計処理をするには文字列型を因子型にしなければいけない

2-2 実演

基本統計量

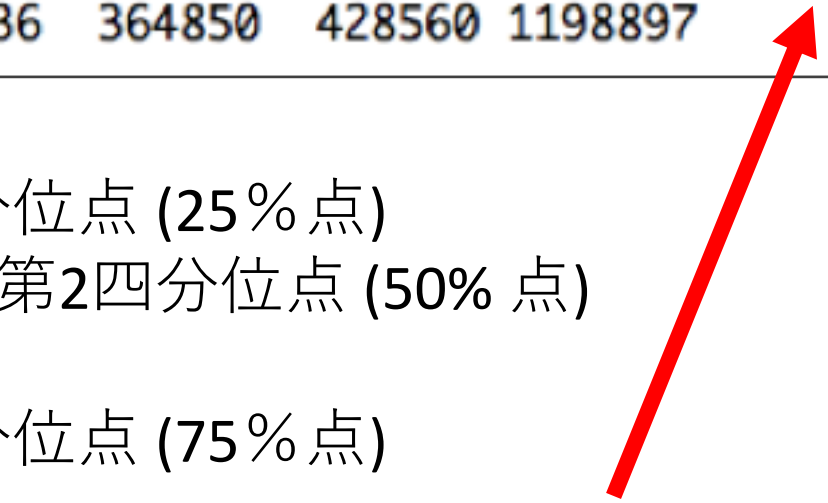
統計量 (statistic)とは、
統計データから計算・要約した数量

基本統計量は、通常広く使用されている
合計、比率、平均、中央値、最頻値、分散、
四分位数 (しぶんいすう) などのことを指す

summary( ) 関数によるデータの要約

```
> summary(Y[Y$BlockID=="D", "Ca"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 121923  272418  347136  364850  428560 1198897      10
```

Min.	最小値
1st Qu.	第1四分位点 (25％点)
Median	中央値 第2四分位点 (50% 点)
Mean	平均値
3rd Qu.	第3四分位点 (75％点)
Max.	最大値
NA’s	**欠損値**



欠損値があるとエラーが出て計算されない関数も多いので注意が必要

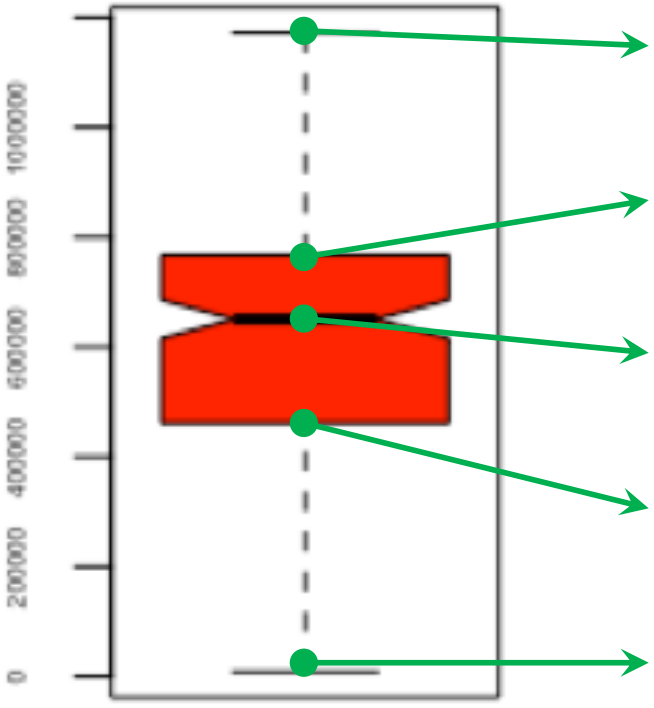
2-3 実演

箱ひげ図 (box plot)

出典: フリー百科事典『ウィキペディア（Wikipedia）』

箱ひげ図（はこひげず、箱髭図、英: box plot、box-and-whisker plot）は、データのばらつきをわかりやすく表現するための統計図である。主に多くの水準からなる分布を視覚的に要約し、比較するために用いる。ジョン・テューキーが1970年代に提唱した。様々な分野で利用されるが、特に品質管理で盛んに用いられる。**箱**（box）と、その両側に出た**ひげ**（whisker）で表現されることからこの名がある[1]。

```
> summary(Y[Y$BlockID=="D", "Ca"])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
 121923  272418  347136  364850  428560 1198897      10
```



Max. 最大値

3rd Qu. 第3四分位点 (75％点)

（しぶんいてん）

Median 中央値 第2四分位点 (50% 点)

1st Qu. 第1四分位点 (25％点)

Min. 最小値

箱ひげ図
(box plot)

Exercise
～ 5 min

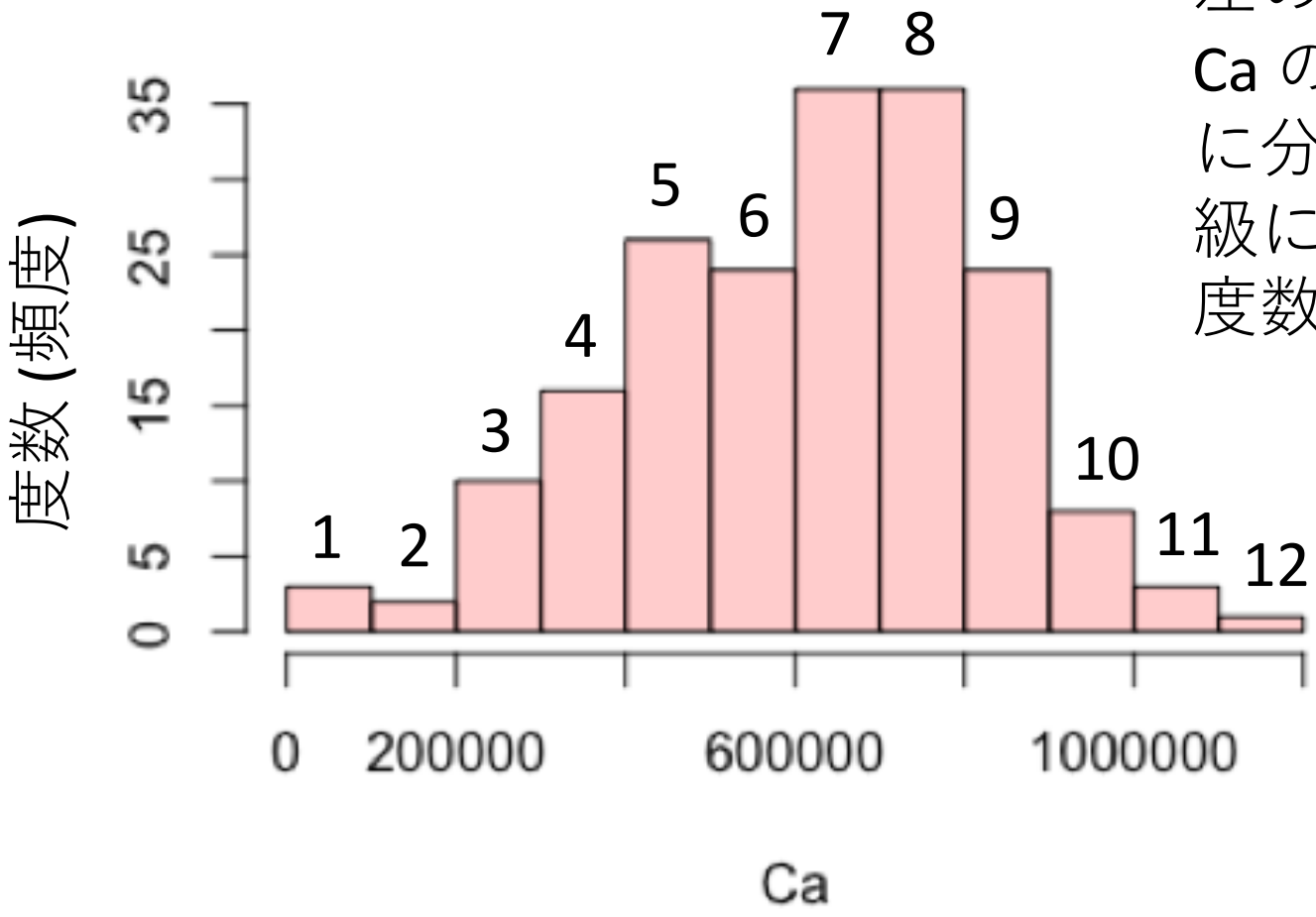
Level 1 : 色やスタイルを変更してみよう
Level 2 : 様々な元素について箱ひげ図を作ってみよう
Level 3 : C と D 以外の比較を行ってみよう

2-4 実演

ヒストグラム (histogram)

出典: フリー百科事典『ウィキペディア（Wikipedia）』

ヒストグラム（英語: histogram[1]）とは、縦軸に度数、横軸に階級をとった統計グラフの一種で、データの分布状況を視覚的に認識するために主に統計学や数学、画像処理等で用いられる。**柱図表**[1]、**度数分布図**、**柱状グラフ**ともいう。



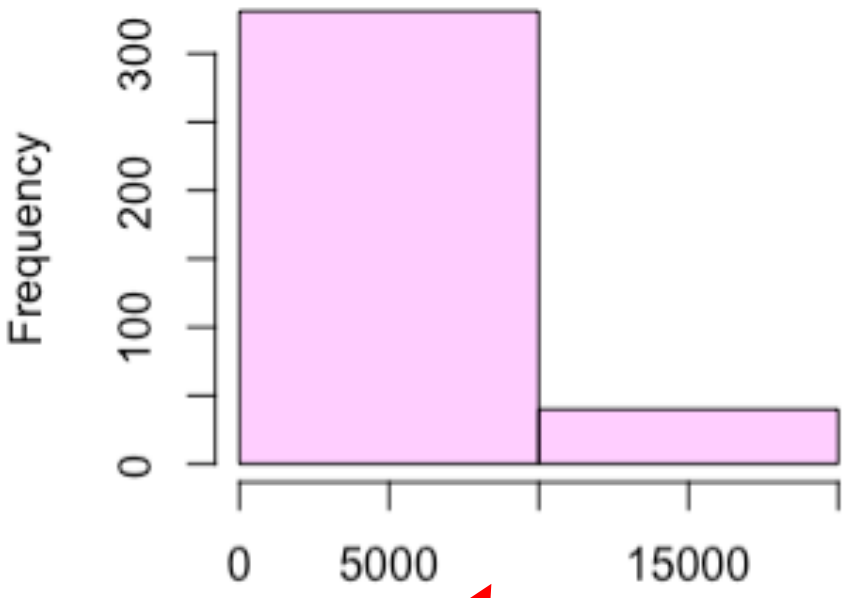
左のヒストグラムでは、Ca の値を 12 の階級 (X 軸) に分けて、それぞれの階級に入るサンプルの数を度数 (Y 軸) にとっている

2-4 実演

ヒストグラム (histogram)

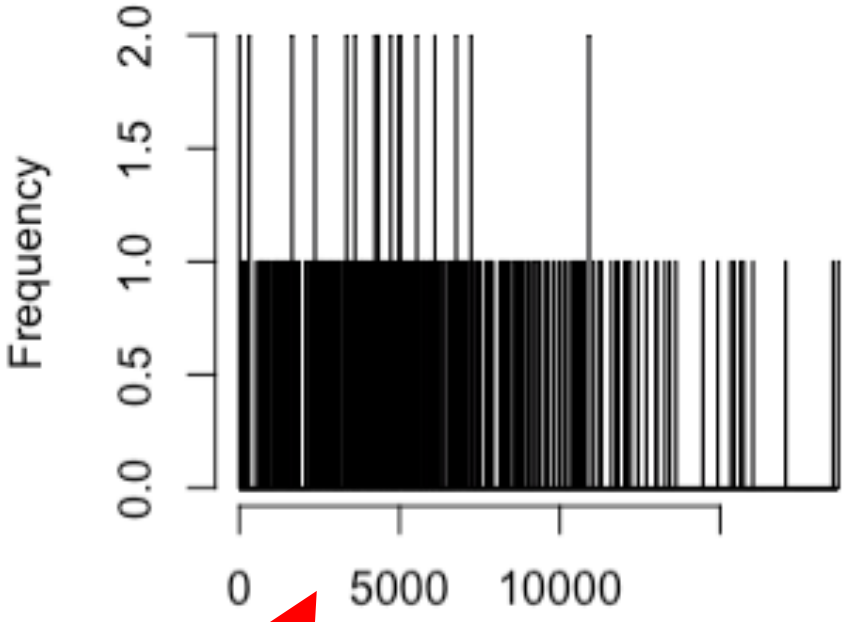
階級幅が小さすぎたり大きすぎたり
するとデータ分布の把握ができない
適切な階級幅を指定することが重要

階級幅 極大



１つの階級にほとんどの
サンプルが含まれている

階級幅 極小



各階級に１つか２つの
サンプルしか含まれていない

ヒストグラム
(histogram)

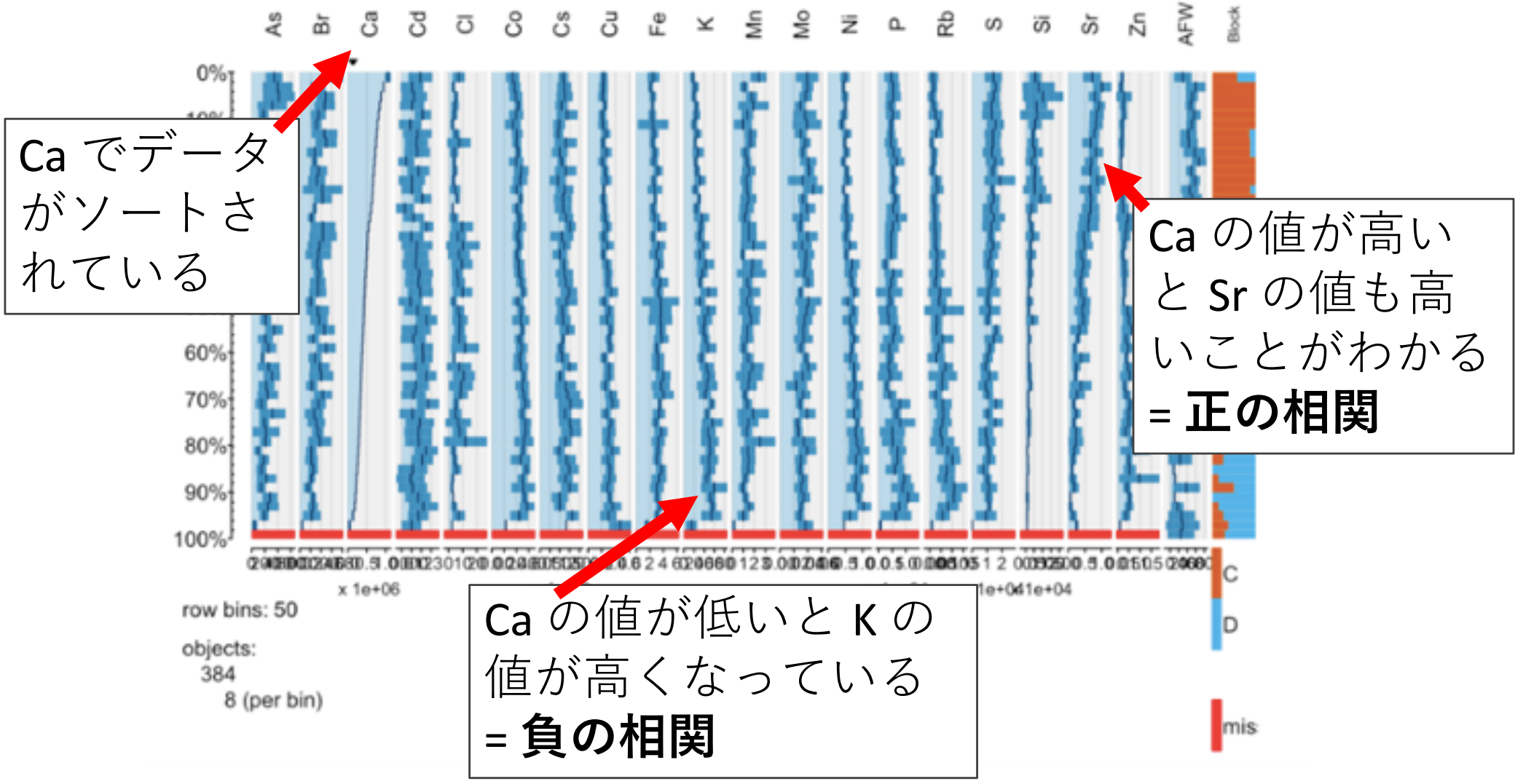
Exercise
～ 5 min

Level 1 : パラメーターを変更してみよう
Level 2 : 様々な元素についてヒストグラムを作ってみよう
Level 3 : C と D 以外の分布も視てみよう

2-5 実演

データの全体を視る

tabplot で変数をソートしてグラフ化し、全体を眺める



Ca でデータがソートされている

Ca の値が高いと Sr の値も高いことがわかる
= **正の相関**

Ca の値が低いと K の値が高くなっている
= **負の相関**

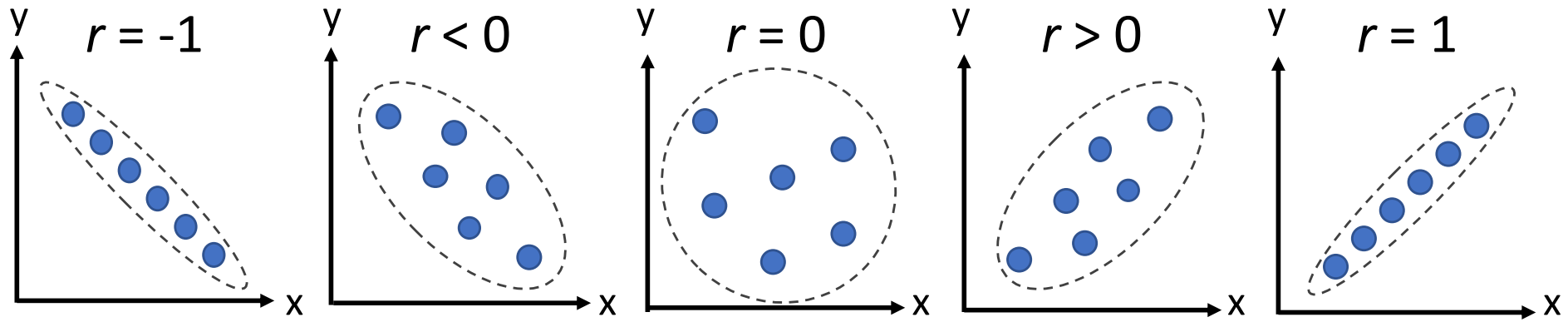
2-6 実演

相関 (correlation)

出典: フリー百科事典『ウィキペディア（Wikipedia）』

相関係数（そうかんけいすう、英: *correlation coefficient*）は、2つの確率変数の間にある線形な関係の強弱を測る指標である[1][2]。相関係数は無次元量で、−1以上1以下の実数に値をとる。相関係数が正のとき確率変数には**正の相関**が、負のとき確率変数には**負の相関**があるという。また相関係数が0のとき確率変数は**無相関**であるという[3][4]。

相関係数 (correlation coefficient : *r*)



x と y の相関係数 (*r*)

$$\frac{\text{(x と y の共分散)}}{\text{(x の標準偏差)} \times \text{(y の標準偏差)}}$$

$r = 0.0 \sim \pm 0.2$	(ほとんど相関がない)
$\pm 0.2 \sim \pm 0.4$	(弱い相関がある)
$\pm 0.4 \sim \pm 0.7$	(相関がある)
$\pm 0.7 \sim \pm 1$	(強い相関がある)

2-6 実演

相関関係の検定

cor.text() 関数

ピアソンの積率相関係数

```
        Pearson's product-moment correlation

data:  Y[Y$BlockID == "C", c("Ca")] and Y[Y$BlockID == "C",
c("Sr")]
t = 9.4901, df = 187, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.4652735 0.6592262
sample estimates:
      cor
0.5701414
```

相関係数 (r)

母集団の
信頼区間

t 値、自由度、**p 値**

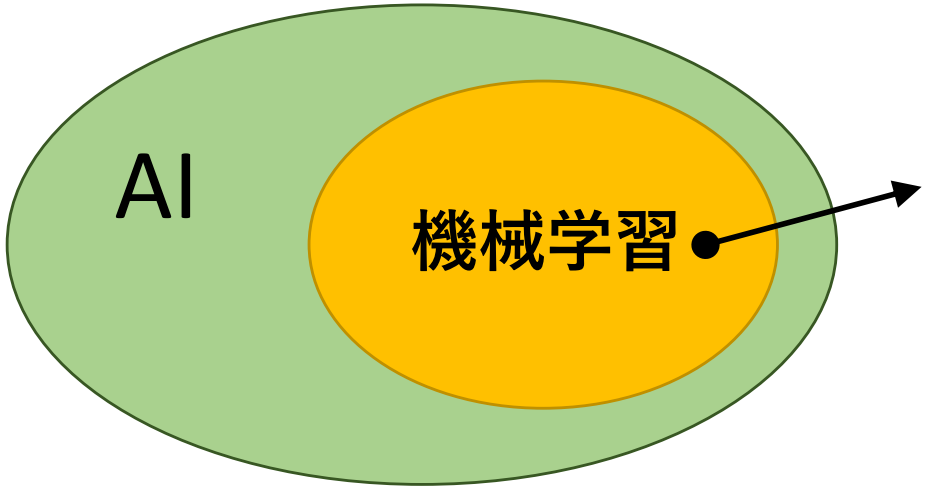
p 値が 0.05 より小さいなら
有意な相関があるとする

(母集団では 相関がないにもかかわらず、検定結果が有意になる
標本が偶然得られた訳ではない)

3 機械学習で分類

Rではじめる機械学習
データサイズを抑えて軽量な環境で攻略法を探る
長橋 賢吾 (著)

機械学習 (machine learning) とは、「大量のデータをもとに、そのデータを処理するプログラムとしてモデルを記述し、そのモデルに基づいて、自動的に意思決定をする」こと



機械学習は AI に内包される

教師あり学習
入力・出力データから予測モデルを開発

教師なし学習
入力データのみからデータをグループ化し解釈

3 機械学習で分類

教師あり学習

分類 (識別) データが所属するグループを**予測**する
(商品を買うか買わないかなど)

回帰 連続した量的データを**予測**する
(広告による商品の売り上げの増加など)

教師なし学習

クラスタリング データに内在するグループ分けを見つけ出す

パターン抽出 データの大部分を表すようなルールを見つけ出す
(コーヒーを買う人はタバコも買う傾向にあるなど)

外れ値検出 異常を見つけ出す

次元削減 効果的な特徴を見つけ出す

3 機械学習で分類

分類や回帰の機械学習アルゴリズム (算法)

最近傍法 (分類)
ナイーブベイズ (分類)
線形回帰 (回帰)
決定木 (分類と回帰)
ニューラルネットワーク (分類と回帰)
サポートベクトルマシン (分類と回帰)
バギング (分類と回帰)
ブースティング (分類と回帰)
ランダムフォレスト (分類と回帰)
, etc.

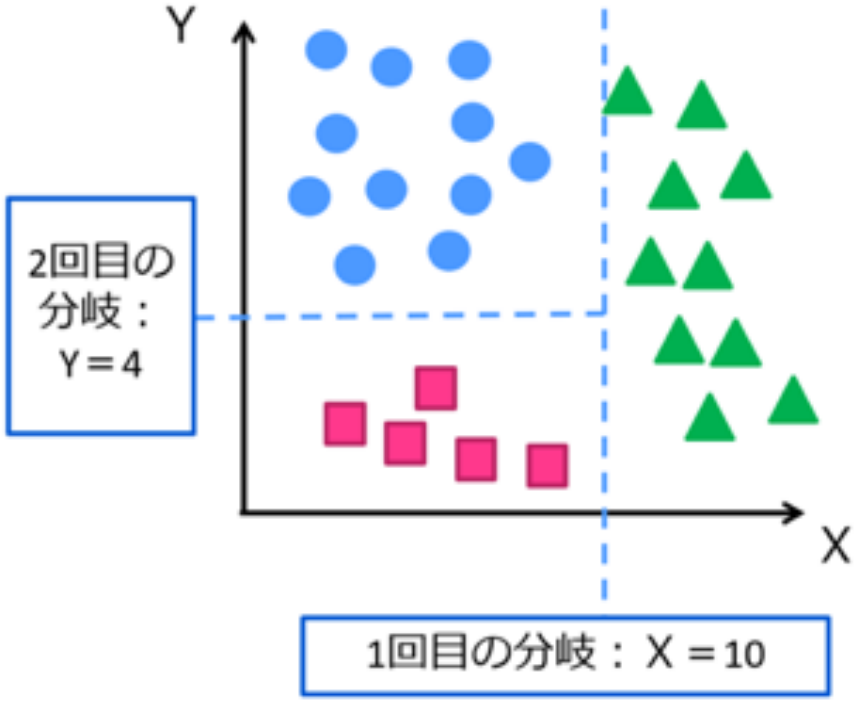
3 機械学習で分類

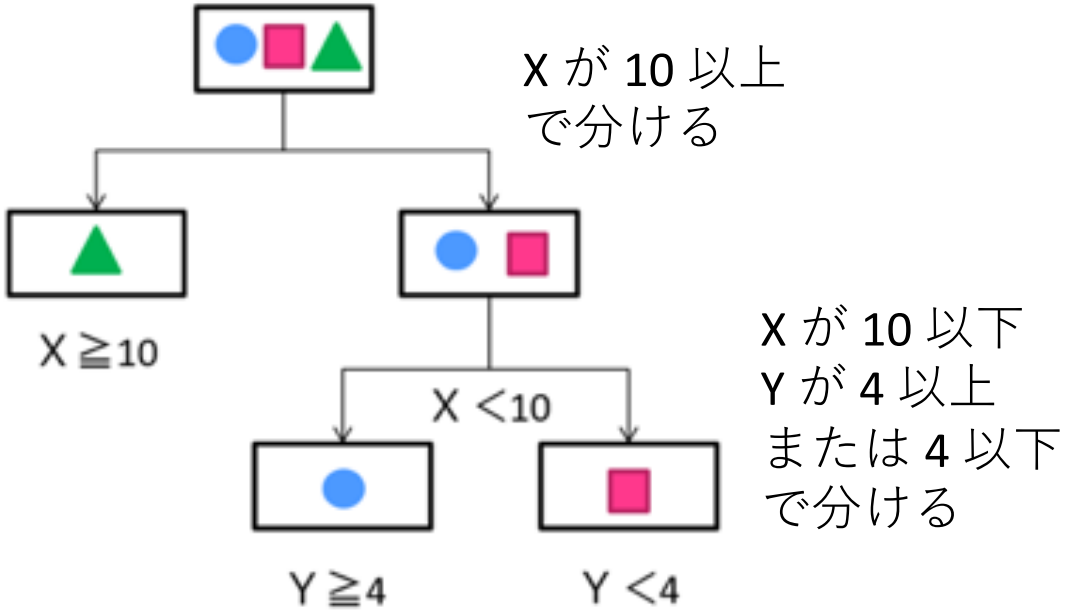
決定木 (Decision tree)

https://toukei-lab.com/決定木
から紹介しています

出典: フリー百科事典『ウィキペディア（Wikipedia）』

決定木（けっていぎ、英: decision tree）は、（リスクマネジメントなどの）決定理論の分野において、 決定を行う為のグラフであり、計画を立案して目標に到達するために用いられる。 決定木は、意志決定を助けることを目的として作られる。 決定木は木構造の特別な形である。





分類に用いる場合、分類木(Classification Tree) と呼び
回帰に用いる場合、回帰木(Regression Tree) と呼ぶ

3 機械学習で分類

ランダムフォレスト (Random forest)

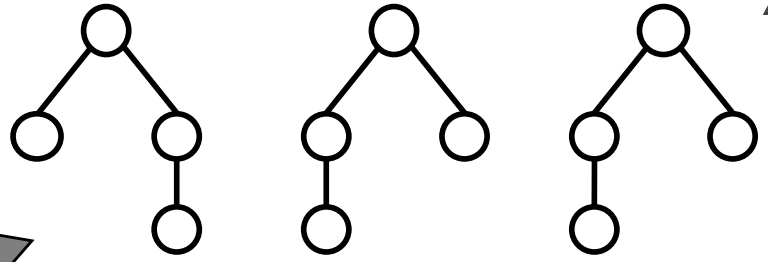
Rによるやさしい
テキストマイニング
：機械学習編
小林雄一郎 (著)

予測モデル

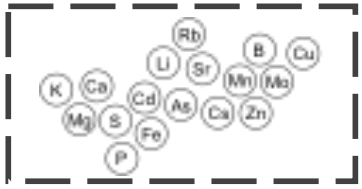
(例) C か D かの分類

ランダムにデータを分割

分類木１　分類木２　分類木３



未知の元素データ



各分類木の予測結果

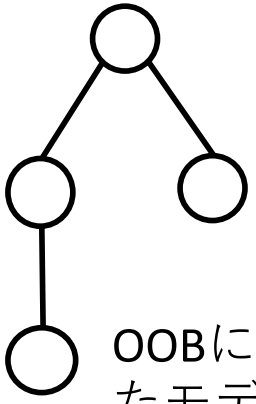
C　C　D

多数決で C と予測

分割データ

学習サンプル　Out-of-bag (OOB)
としてさらに分割

各分類木



説明変数をランダムに選択
しうまくC とDを分割できる変数を選んで分枝する
(ジニ不純度で判断)

OOBに対して学習用データで構築したモデルを当てはめた時の推定誤差から、**説明変数の重要度を計算**

3-1 実演

機械学習で分類

randomForest() 関数で 予測モデル を作る

予測に用いる変数を "y ~ x1 + x2 " のように記述

y は目的変数 (従属変数)、x は説明変数 (独立変数)と呼ばれる

```
randomForest(BlockID ~
        As+Br+Ca+Cd+Cl+Co+Cs+Cu+Fe+K+Mn+Mo+Ni+P+Rb+S+Si+Sr+Zn,
        data = X2,
        proximity = TRUE, # 個々のデータの近接性を計算、MDS plot での結果の可視化に使用
        importance = TRUE # 特徴量加工による重要度も計算・出力
        )
```

x が行列なら、
randomforest(x = 行列,
y = ベクトル,
proximity = TRUE,
importance =TRUE
) としても良い

近接性 (データの類似度)
2 つのデータに対して、ランダムフォレストの各決定木の同じ終端ノードに落ちた割合

3-1 実演

機械学習で分類
予測モデルの評価

```
Call:
 randomForest(formula = BlockID ~ As + Br + Ca + Cd + Cl + Co +      Cs + Cu +
Ni + P + Rb + S + Si + Sr +      Zn, data = X2, proximity = TRUE, importance =
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 2.43%
Confusion matrix:
    C   D class.error
C 184   5  0.02645503
D   4 178  0.02197802
```

モデルの種類
classification = 分類

分類木の数

決定木の分割の時に選択する説明変数の数

Out of bag (OOB)
でのエラー率

予測の不正解率 = 2.43%、正解率 = 97.57%

モデル作成に使った全データでの不正解率、
9つのサンプルについてうまく分類できなかった

3-2 機械学習で分類

https://www1.doshisha.ac.jp/~mjin/R/Chap_27/27.html
[連載]フリーソフトによるデータ解析・マイニング第27回から改変して紹介しています

多次元尺度構成法 (MultiDimensional Scaling: MDS)

データの個体間の類似度や距離を求めてそれを 2 ～ 3 次元にプロットして**データの構造やパターン形成などを把握**する手法

近畿地方の地図

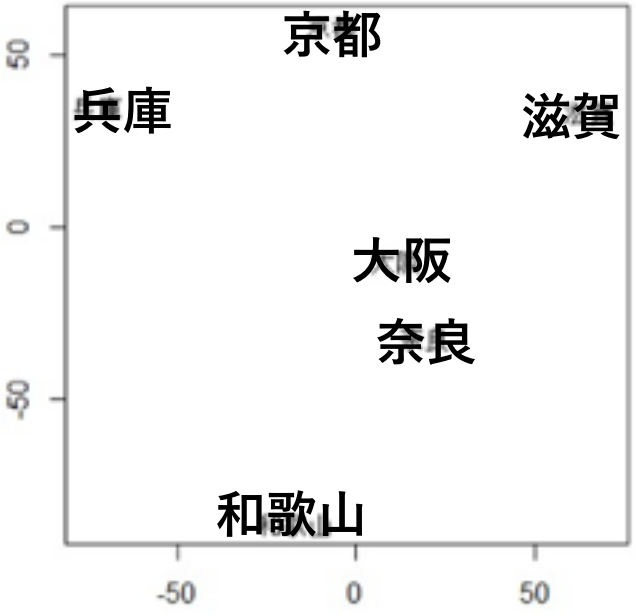
兵庫県からの距離行列 (km)

	兵庫	和歌山	大阪	奈良	滋賀	京都
兵庫	0	134	85	116	118	60
和歌山	134	0	68	66	145	141
大阪	85	68	0	32	83	75
奈良	116	66	32	0	79	95
滋賀	118	145	83	79	0	63
京都	60	141	75	95	63	0

距離行列から求めた 2 次元座標値

	横軸	縦軸
兵庫	-71.9	35.1
和歌山	-16.8	-85.9
滋賀	65.0	33.8
京都	-6.7	58.6
奈良	19.3	-32.1
大阪	11.0	-9.5

2 次元座標値の散布図



3-2 機械学習で分類

近接性 (データの類似度)

```
> ionome.rf.class$proximity[1:10,1:5]
           1         2         3         4         5
1  1.0000000 0.7647059 0.4558824 0.7066667 0.2089552
2  0.7647059 1.0000000 0.4590164 0.6794872 0.1764706
3  0.4558824 0.4590164 1.0000000 0.5555556 0.1132075
4  0.7066667 0.6794872 0.5555556 1.0000000 0.1194030
5  0.2089552 0.1764706 0.1132075 0.1194030 1.0000000
6  0.3424658 0.3157895 0.1764706 0.3389831 0.2727273
7  0.6197183 0.6538462 0.2758621 0.5312500 0.2372881
8  0.4098361 0.4035088 0.2280702 0.2985075 0.1904762
9  0.8250000 0.8648649 0.4603175 0.6875000 0.1969697
10 0.8148148 0.8135593 0.6140351 0.5915493 0.2205882
```

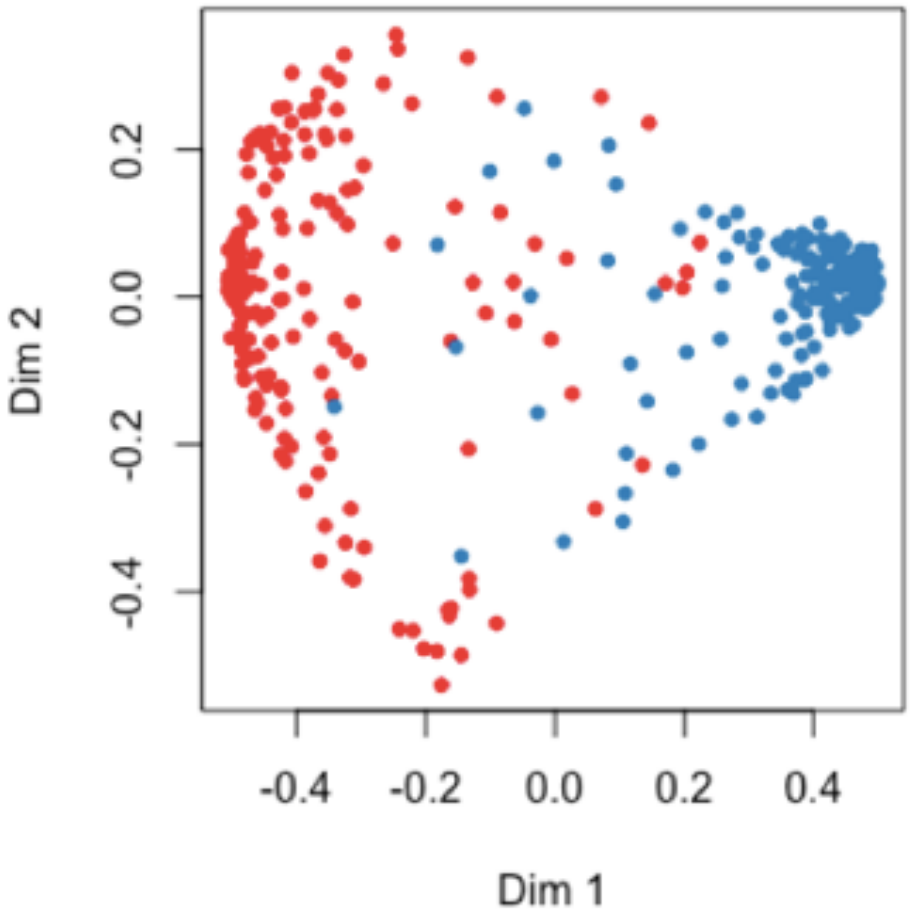
randomForest() 関数により近接性の行列が得られており、MDS plot の作成にはこれが使われている

3-2 実演

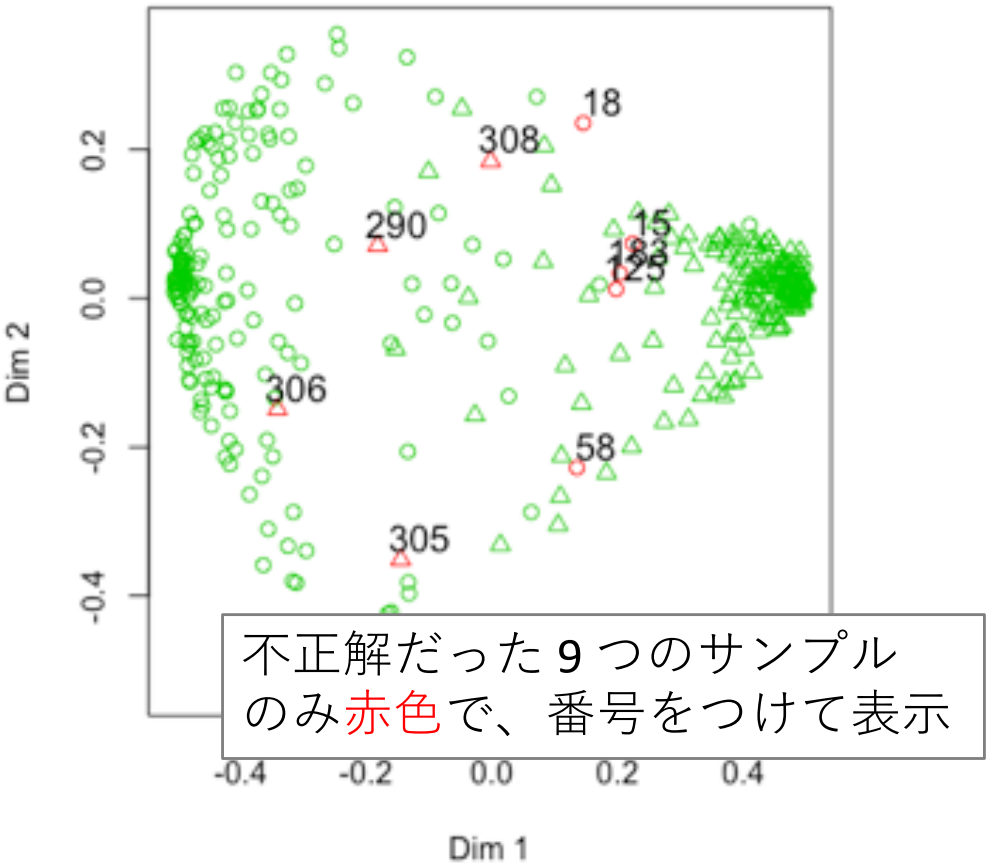
機械学習で分類

MDS plot で誤分類されたサンプルを視る

C のサンプル
D のサンプル



◯ C のサンプル
△ D のサンプル



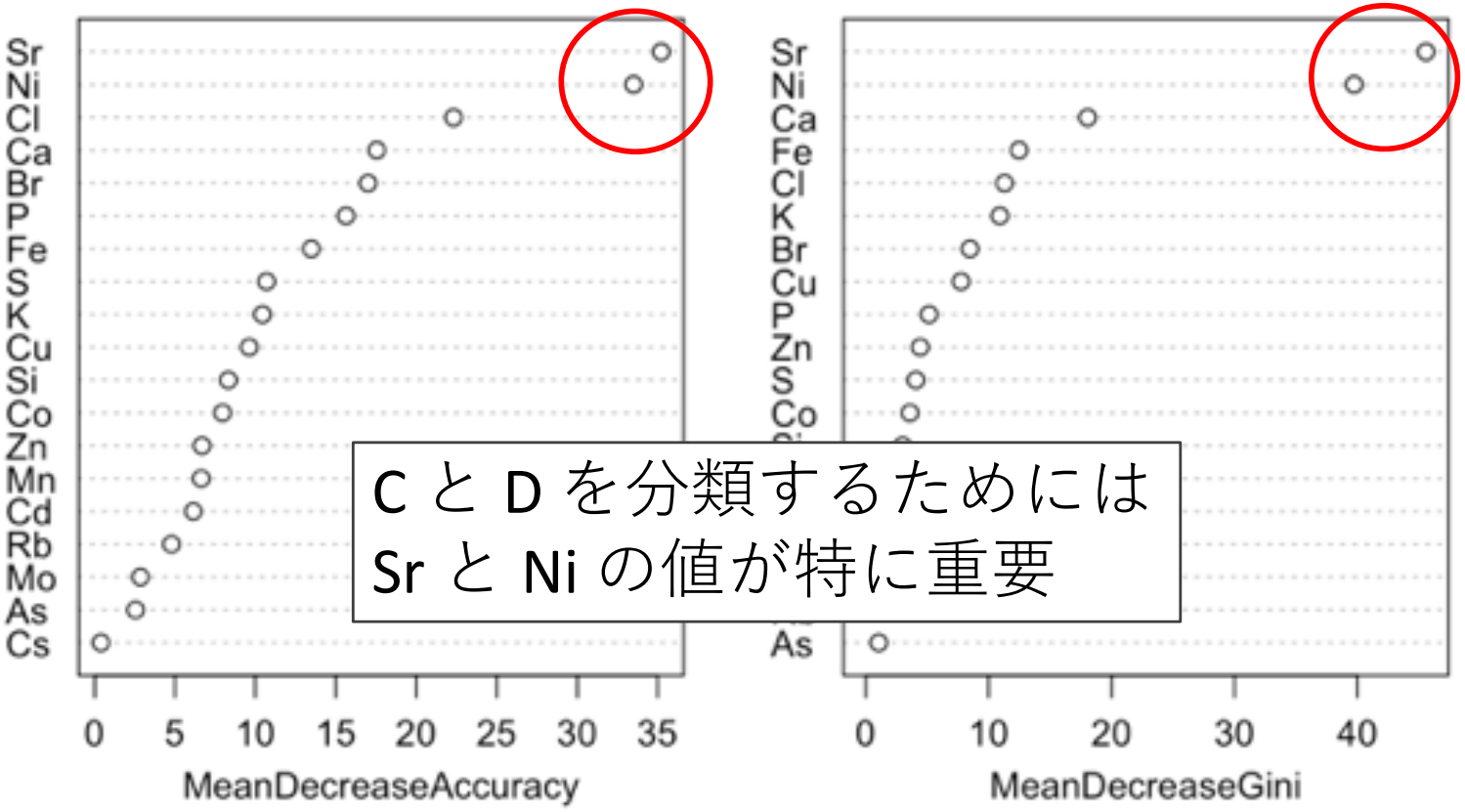
不正解だった 9 つのサンプル
のみ赤色で、番号をつけて表示

3-3 実演

機械学習で分類

説明変数の重要度を視る

特徴量加工による重要度　　ジニ係数による重要度

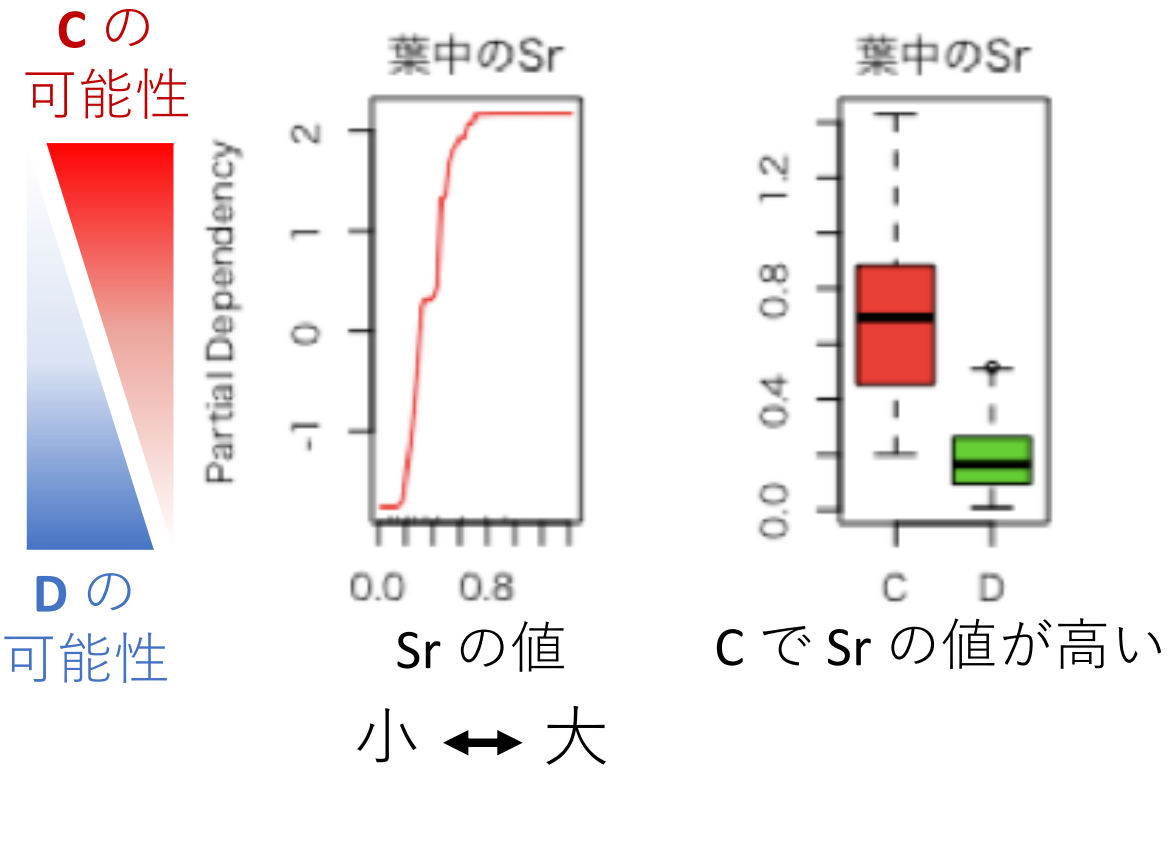


3-4 実演

機械学習で分類

分類の結果を partialPlot (部分従属プロット) で視る

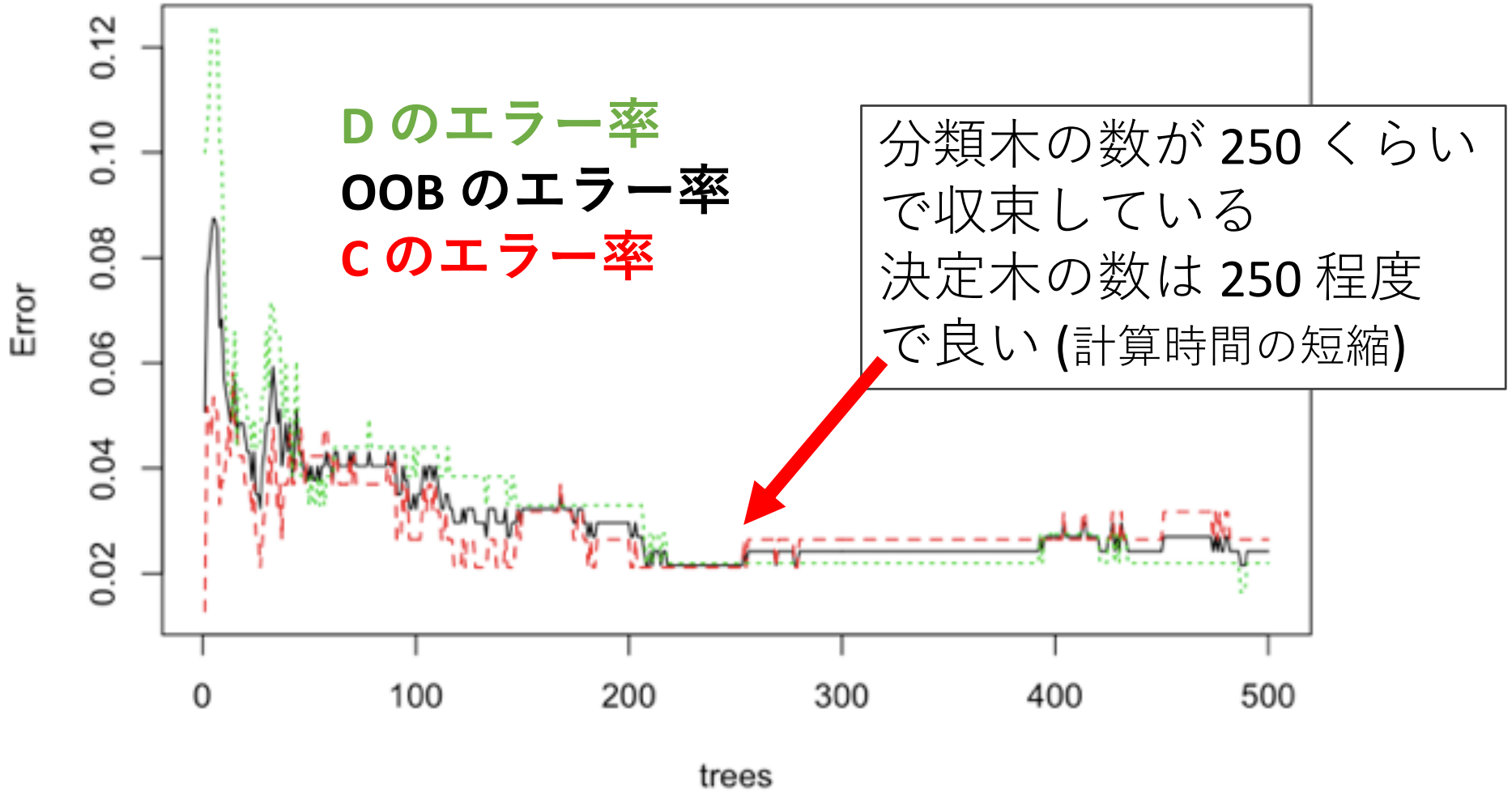
重要な説明変数がどのように分類に寄与しているかを視る



3-5 実演

機械学習で分類

学習の収束状況を視る



3-6 機械学習で分類

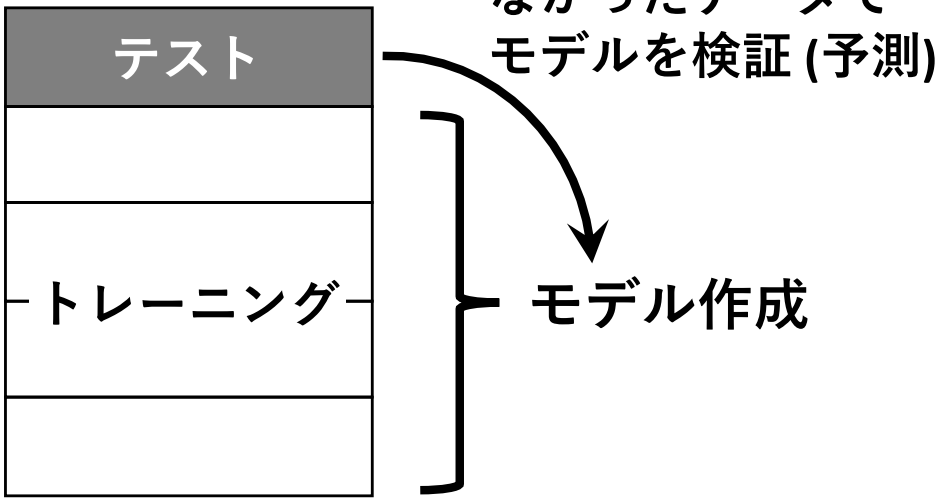
交差検定 (Cross-validation)

出典: フリー百科事典『ウィキペディア（Wikipedia）』

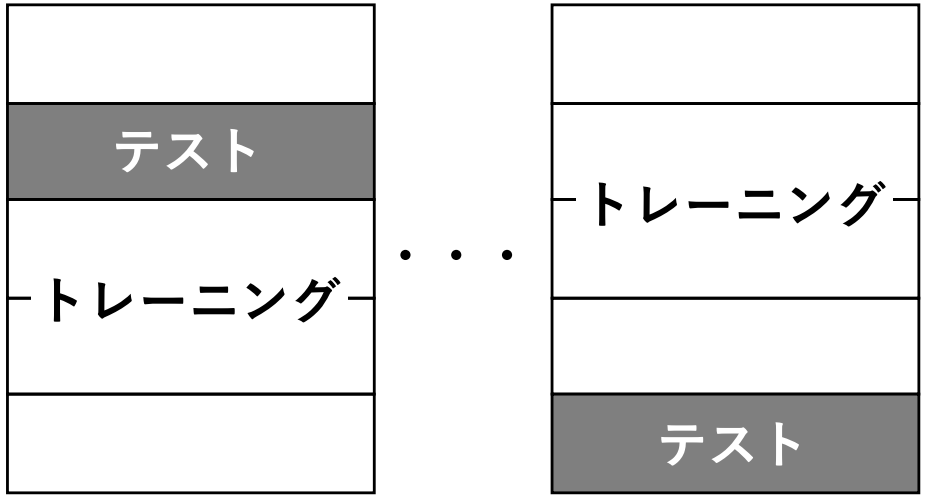
交差検証（交差確認）[1]（こうさけんしょう、英: **Cross-validation**）とは、統計学において標本データを分割し、その一部をまず解析して、残る部分でその解析のテストを行い、解析自身の妥当性の検証・確認に当てる手法を指す[2] [3] [4]。データの解析（および導出された推定・統計的予測）がどれだけ本当に母集団に対処できるかを良い近似で検証・確認するための手法である。

最初に解析するデータを「訓練事例集合（training set）」などと呼び、他のデータを「テスト事例集合（testing set、テストデータ）」などと呼ぶ。

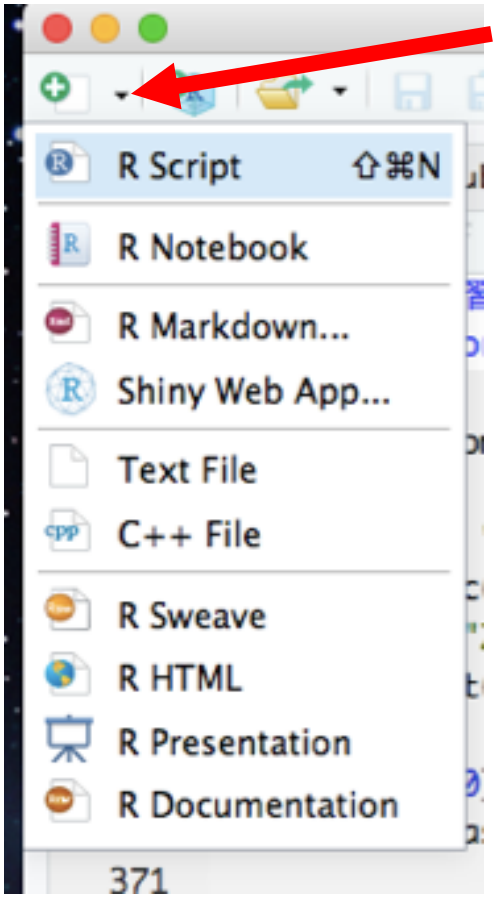
全データを
テストデータと
トレーングセット
に分ける
(ここでは 5 分割)



分割分繰り返し、結果を統合することでモデルの精度を検証する

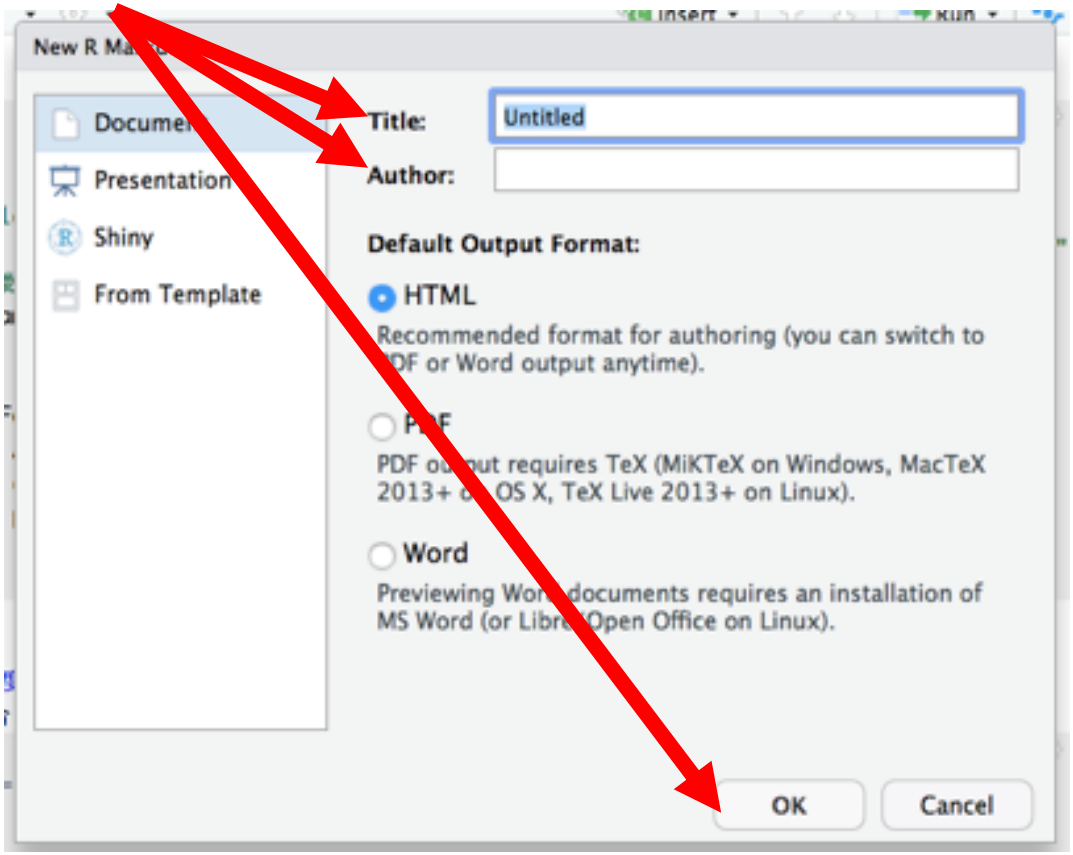


R Markdown でレポート作成



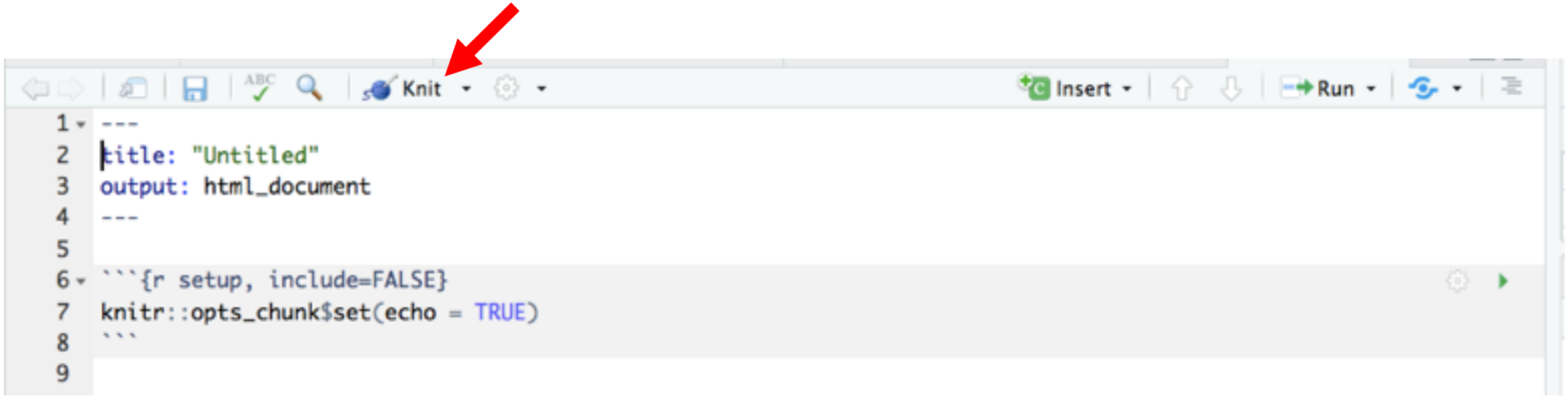
①メニューの左上にある新規ファイル作成ボタンから「R Markdown」をクリック

②タイトルと名前を記入して「OK」をクリック



R Markdown でレポート作成

③「Knit」をクリックしてRmd ファイルを
html ドキュメントに変換 (レンダリング)



④ワーキングディレクトリ内にできた html ファイルを開く

R Markdown でレポート作成

Rmdファイル

```
---
title: "Untitled"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

R Markdown

This is an R Markdown document. Markdowr
documents. For more details on using R M

When you click the **Knit** button a doc
output of any embedded R code chunks wit

```{r cars}
summary(cars)
```

```

## レンダリングされたファイル

# Untitled

## R Markdown

This is an R Markdown document. Markdown is a simple for
details on using R Markdown see http://rmarkdown.rstudio.

When you click the **Knit** button a document will be generate
chunks within the document. You can embed an R code chu

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

R Markdown でレポート作成

Rmdファイル

R チャンク

```
15
16 ```{r cars}
17 summary(cars)
18 ```
19
```

バッククォート (Shift + @)
「```」に挟まれた部分を
R チャンクと呼ぶ

R チャンクの基本型

```
```{r}
R のコード
```
```

レンダリングされたファイル

```
summary(cars)
```

```
##      speed           dist
##  Min.   : 4.0   Min.   :  2.00
##  1st Qu.:12.0   1st Qu.: 26.00
##  Median :15.0   Median : 36.00
##  Mean   :15.4   Mean   : 42.98
##  3rd Qu.:19.0   3rd Qu.: 56.00
##  Max.   :25.0   Max.   :120.00
```

レンダリング後、Rチャンク内のコードの結果が自動的に差し込まれる

R チャンク内に R のコードを記述し、R チャンク外に文章を書くことでドキュメントを作成

R Markdown でレポート作成

講義で使用したR コードを使ってR による機械学習のレポートを作成してみよう

Rユーザのための RStudio[実践]入門
–tidyverseによるモダンな分析フローの世界–
松村 優哉 (著),
湯谷 啓明 (著),
紀ノ定 保礼 (著),
前田 和寛 (著)

ドキュメント・プレゼンテーション生成
シリーズUseful R / 金 明哲編 第 第 9 巻
高橋 康介 (著)