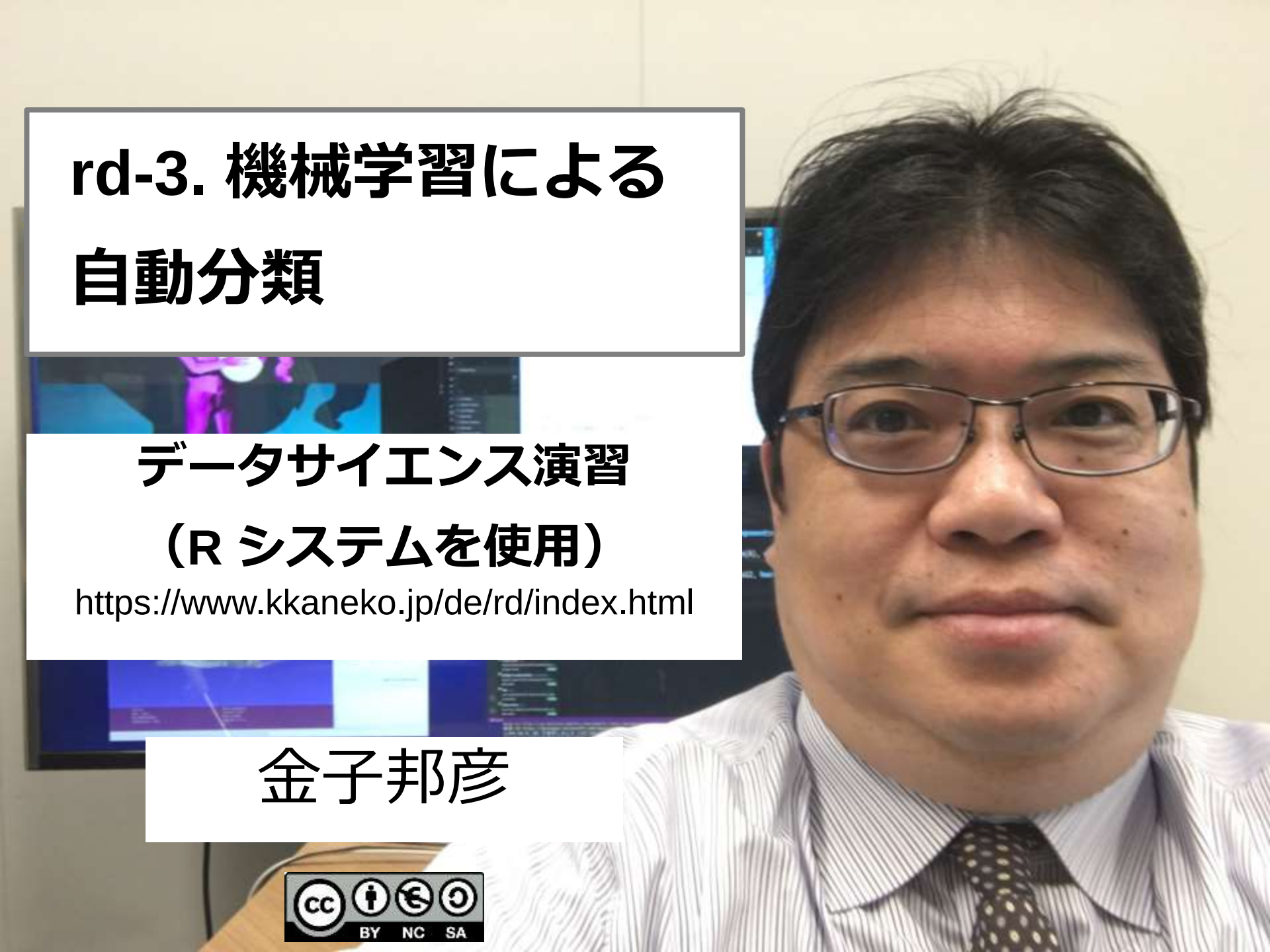
rd-3. 機械学習による自動分類

データサイエンス演習（R システムを使用）

https://www.kkaneko.jp/de/rd/index.html



金子邦彦

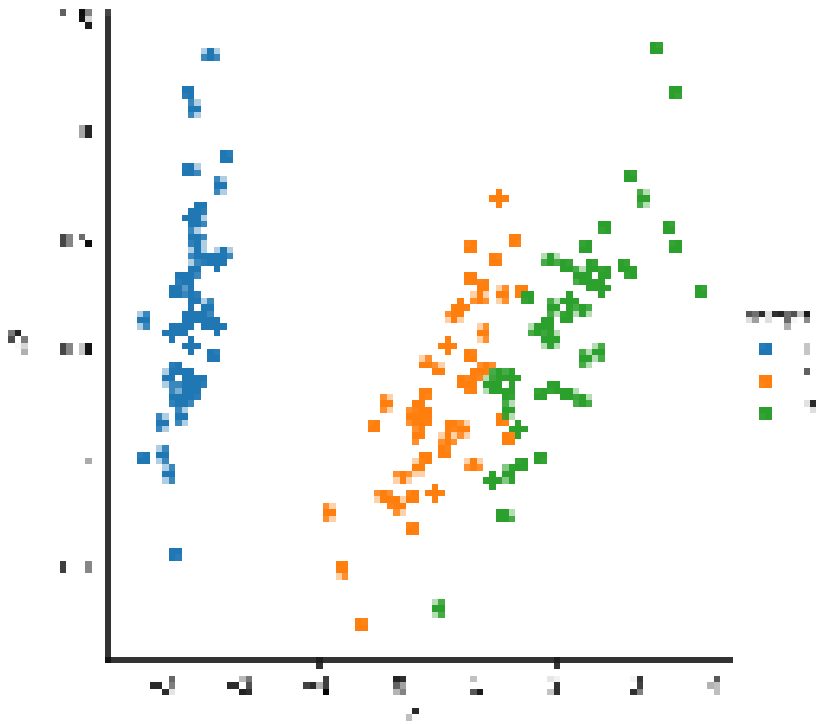
機械学習

- **機械学習**とは、
  与えられたデータ（教師データ）を使い，
  未知のデータに対しても**当てはまる**
  パターンや規則を，コンピュータが抽出
  すること

機械学習の用途

- 未知のデータの分類
- 予測
- 幅広い応用：画像認識，音声認識，自然言語処理，データ分析

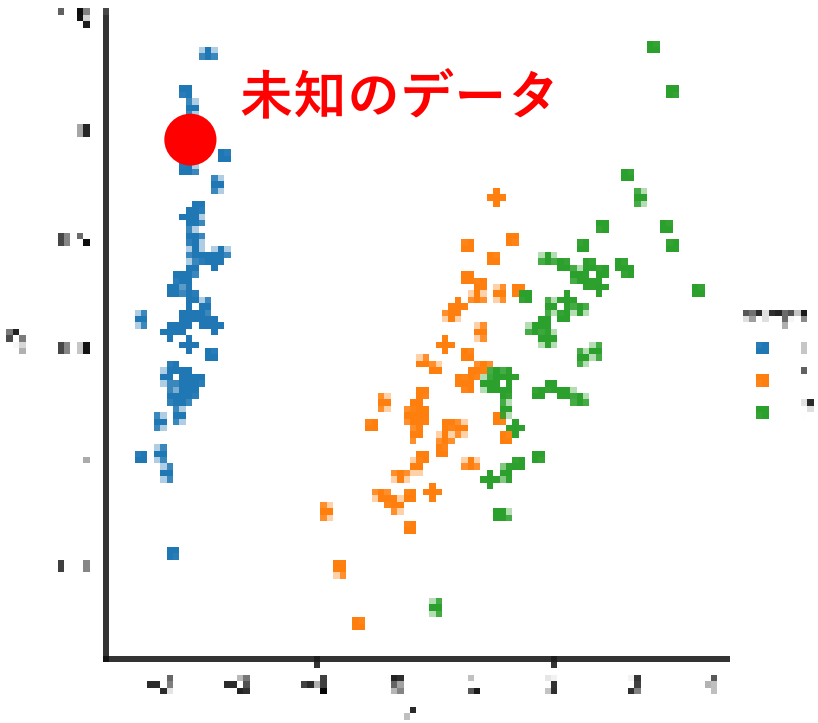
教師データの例



Iris データセット
・3種，150のアヤメの花びらのデータ
※ 右図は，主成分分析の結果のプロット

- **教師データ**は，**多数のデータの集まり**
- 上の図では，点１つで，１つのデータ

自動分類

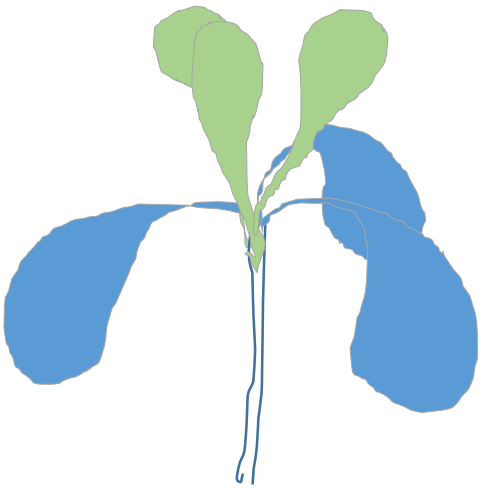


- 新しいデータ（**未知のデータ**）が
  あるとき，花の種類は何でありそうか

教師データの利用により，**未知のデータ**についても見通しを立てることが可能に

3-1. Iris データセット

アヤメ属 (Iris)



- 多年草
- 世界に 150種. 日本に 9種.
- 花被片は 6個
- 外花被片（がいかひへん） Sepal
  3個（大型で下に垂れる）
- 内花被片（ないかひへん） Petal
  3個（直立する）

Iris データセット

```
> iris
   Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1           5.1         3.5          1.4         0.2     setosa
2           4.9         3.0          1.4         0.2     setosa
3           4.7         3.2          1.3         0.2     setosa
4           4.6         3.1          1.5         0.2     setosa
5           5.0         3.6          1.4         0.2     setosa
6           5.4         3.9          1.7         0.4     setosa
7           4.6         3.4          1.4         0.3     setosa
8           5.0         3.4          1.5         0.2     setosa
9           4.4         2.9          1.4         0.2     setosa
10          4.9         3.1          1.5         0.1     setosa
11          5.4         3.7          1.5         0.2     setosa
12          4.8         3.4          1.6         0.2     setosa
```

Iris データセットは，Rシステムの中に組み込み済み

- 3種のアヤメの外花被辺、内花被片の幅と長さを計測したデータセット

  Iris setosa

  Iris versicolor

  Iris virginica

- データ数は 50 × 3
- 作成者：Ronald Fisher
- 作成年：1936

R システム での Iris データセットの表示

コンソールで次のコマンドを実行

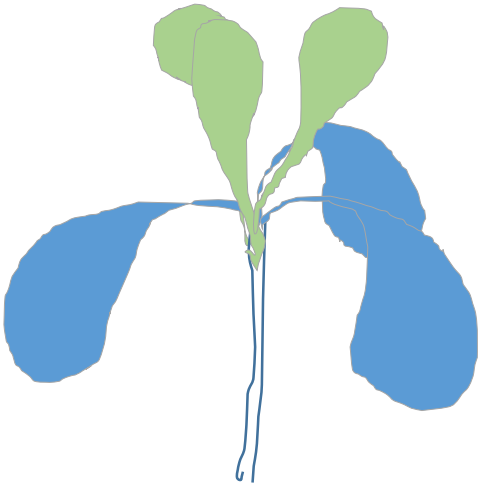
```
iris
```

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
```

コンソール画面をスクロール.

Sepal.Length, Sepal.Width, Petal.Length. Petal.Width, Species の 5属性がある

内花被片（ないかひへん）
Petal



外花被片（がいかひへん）
Sepal

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
6            5.4         3.9          1.7         0.4     setosa
7            4.6         3.4          1.4         0.3     setosa
8            5.0         3.4          1.5         0.2     setosa
9            4.4         2.9          1.4         0.2     setosa
10           4.9         3.1          1.5         0.1     setosa
11           5.4         3.7          1.5         0.2     setosa
12           4.8         3.4          1.6         0.2     setosa
13           4.8         3.0          1.4         0.1     setosa
14           4.3         3.0          1.1         0.1     setosa
15           5.8         4.0          1.2         0.2     setosa
```

外花被片の長さと幅

内花被片の長さと幅

花の種類

R システムでの実行手順

① 必要なパッケージのインストール

コンソールで次のコマンドを実行（コピペ）

```
install.packages("ggplot2")
install.packages("dplyr")
install.packages("klaR")
```

```
> install.packages("ggplot2")
WARNING: Rtools is required to build R pac
kages but is not currently installed. Plea
se download and install the appropriate ve
rsion of Rtools before proceeding:
```

以下省略

Iris データセットの散布図

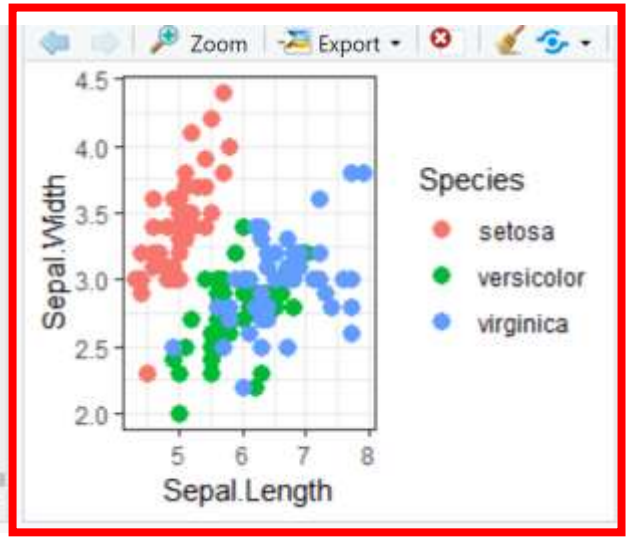
② Sepal.Length, Sepal.Width で**散布図**の作成.

花の種類で色を変える

コンソールで次のコマンドを実行（コピペ）

```
library(ggplot2)
ggplot(iris, aes(x=Sepal.Length)) +
  geom_point( aes(y=Sepal.Width, colour=Species), size=3 ) +
  theme_bw()
```

```
https://cran.rstudio.com/bin/windows/Rtools/
Installing package into 'F:/me/Documents/R/win-library/3.5'
(as 'lib' is unspecified)
Warning in install.packages :
  package 'ggplot2' is in use and will not be installed
> library(ggplot2)
> ggplot(iris, aes(x=Sepal.Length)) +
+     geom_point( aes(y=Sepal.Width, colour=Species), size=3 ) +
+     theme_bw()
>
>
```



散布図が表示されるので確認

Iris データセットの散布図

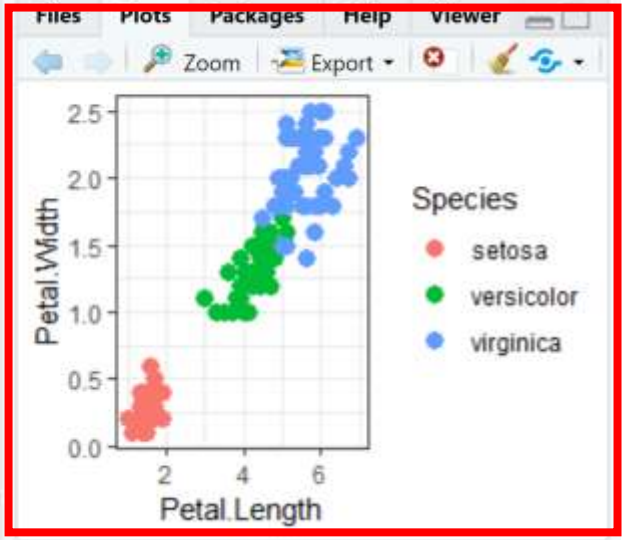
③ Petal.Length, Petal.Width で**散布図**の作成.

花の種類で色を変える

コンソールで次のコマンドを実行（コピペ）

```
library(ggplot2)
ggplot(iris, aes(x=Petal.Length)) +
  geom_point( aes(y=Petal.Width, colour=Species), size=3 ) +
  theme_bw()
```

```
(as "lib" is unspecified)
Warning in install.packages :
  package 'ggplot2' is in use and will not
 be installed
> library(ggplot2)
> ggplot(iris, aes(x=Sepal.Length)) +
+     geom_point( aes(y=Sepal.Width, colou
r=Species), size=3 ) +
+     theme_bw()
>
> library(ggplot2)
> ggplot(iris, aes(x=Petal.Length)) +
+     geom_point( aes(y=Petal.Width, colou
r=Species), size=3 ) +
+     theme_bw()
>
>
```



散布図が表示されるので確認

3-2. 学習

自動分類のための学習

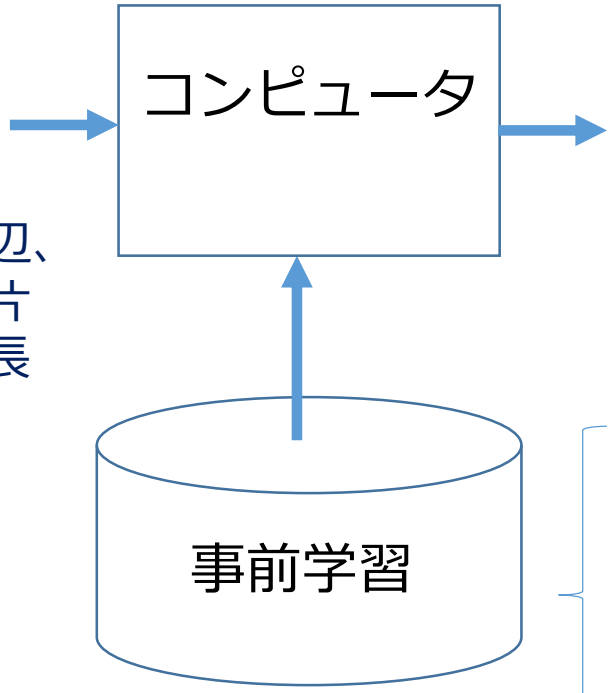
- **属性データ**と**その種類に関するデータ（ラベルなどという）**を使って、**学習**

- 教師あり学習（Supervised Learning）ともいう

自動分類のための学習

新しい数値の並び

外花被辺、内花被片の幅と長さ



ラベルを予測,

ラベルは，Iris setosa, Iris versicolor, Iris virginica のいずれか

Iris setosa	Iris versicolor	Iris virginica
5.1 3.5 1.4 0.2	7.0 3.2 4.7 1.4	6.3 3.3 6.0 2.5
4.9 3.0 1.4 0.2	6.4 3.2 4.5 1.5	5.8 2.7 5.1 1.9
4.7 3.2 1.3 0.2	6.9 3.1 4.9 1.5	7.1 3.0 5.9 2.1
4.6 3.1 1.5 0.2	5.5 2.3 4.0 1.3	6.3 2.9 5.6 1.8
5.0 3.6 1.4 0.2	6.5 2.8 4.6 1.5	6.5 3.0 5.8 2.2

外花被辺、内花被片の幅と長さ

自動分類のための学習

学習のデータセットは次の形をしている

数値の並び　＋　ラベル

教師あり学習 (supervised learning) のデータセットの例

- Iris データセットは，3種のアヤメの外花被辺、内花被片の幅と長さを計測したデータセット

5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	satosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
…				

7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
5.5	2.3	4.0	1.3	versicolor
6.5	2.8	4.6	1.5	versicolor
…				

6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica
6.3	2.9	5.6	1.8	virginica
6.5	3.0	5.8	2.2	virginica
…				

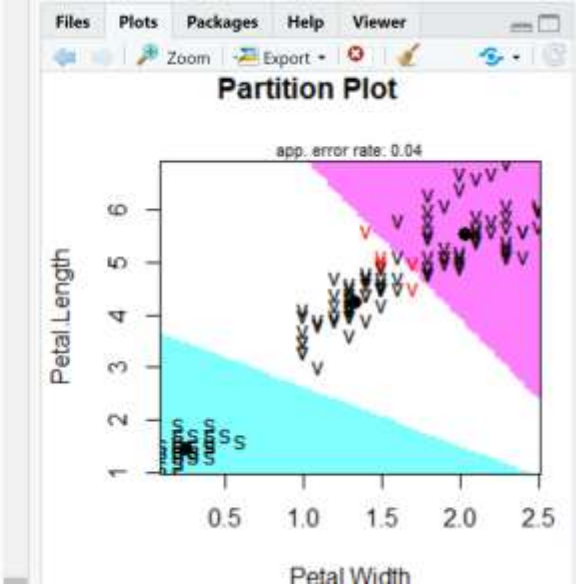
数値の並び　＋　**ラベル**（花の種類）

LDA 法のプログラム例

Rstudio のコンソールで次のコマンドを実行

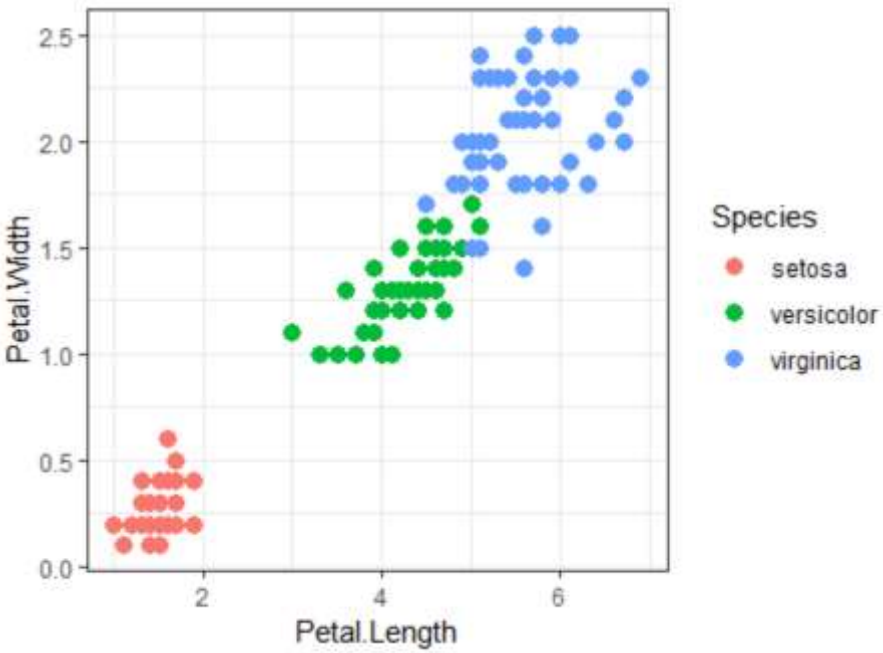
```
library(dplyr)
library(klaR)
d <- tbl_df(iris[c(3,4,5)])
partimat(Species~., data=d, method="lda")
```

```
以下のオブジェクトは 'package:ggplot2' からマスクされています:

    vars

以下のオブジェクトは 'package:stats' からマスクされています:

    filter, lag

以下のオブジェクトは 'package:base' からマスクされています:

    intersect, setdiff, setequal, union

> library(klaR)
要求されたパッケージ MASS をロード中です

次のパッケージを付け加えます: 'MASS'

以下のオブジェクトは 'package:dplyr' からマスクされています:

    select

> d <- tbl_df(iris[c(3,4,5)])
> partimat(Species~., data=d, method="lda")
>
```

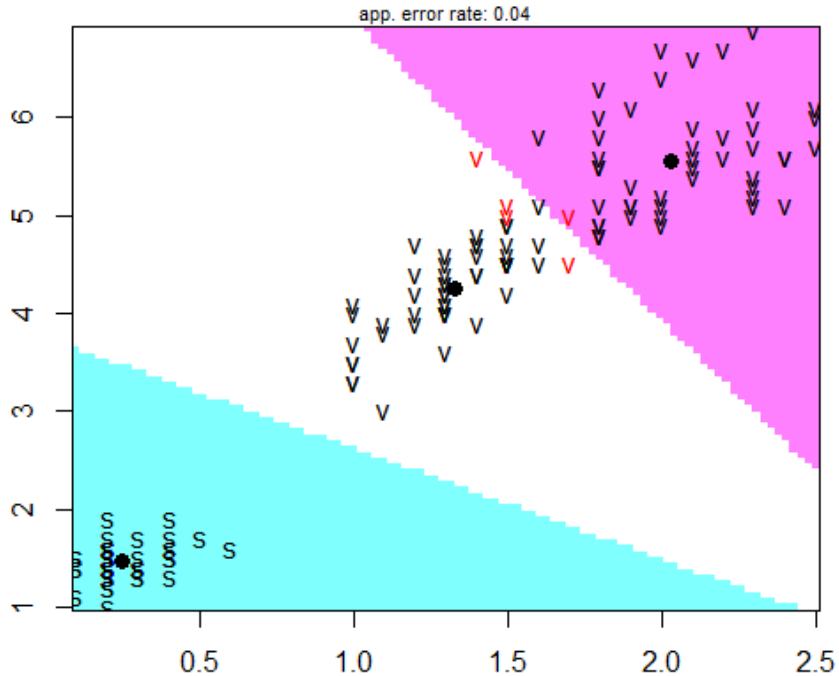


赤、白、水色の
パーティション

LDA 法は、教師あり学習の１手法



元データ



空間が区分けされた.

学習のデータセット

- **数値の並び**
  内花被片の幅と高さのデータ
- **ラベル**
  花の種類のデータ

新しい数値（内花被片の幅と高さ）が得られたとき、花の種類を予測できる