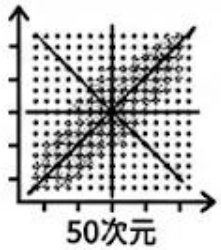
pd-2. 主成分分析，次元削減

（Python によるデータサイエンス演習）

URL: https://www.kkaneko.jp/ai/pd/index.html

金子邦彦

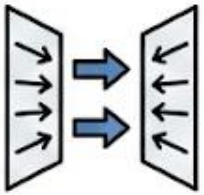
多次元データの課題



- 高次元で全体像の把握が困難
- 計算コストが増大

STEP 1

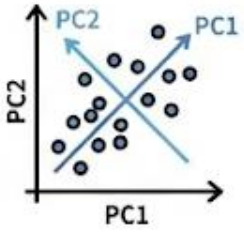
次元削減の概要と目的



- 可視化
- ノイズ除去・特徴抽出
- 計算効率化

STEP 2

主成分分析（PCA）の原理



- 分散最大化による主軸の決定
- 上位主軸への投影で要約

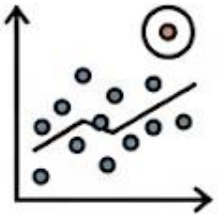
STEP 3

実践演習（Irisデータ）

- データ準備（スケーリングのプランクの重要性）
- PCA実行
- 結果確認（2次元プロット）

STEP 4

応用：外れ値検出



- 外れ値の識別
- ロバスト主成分分析

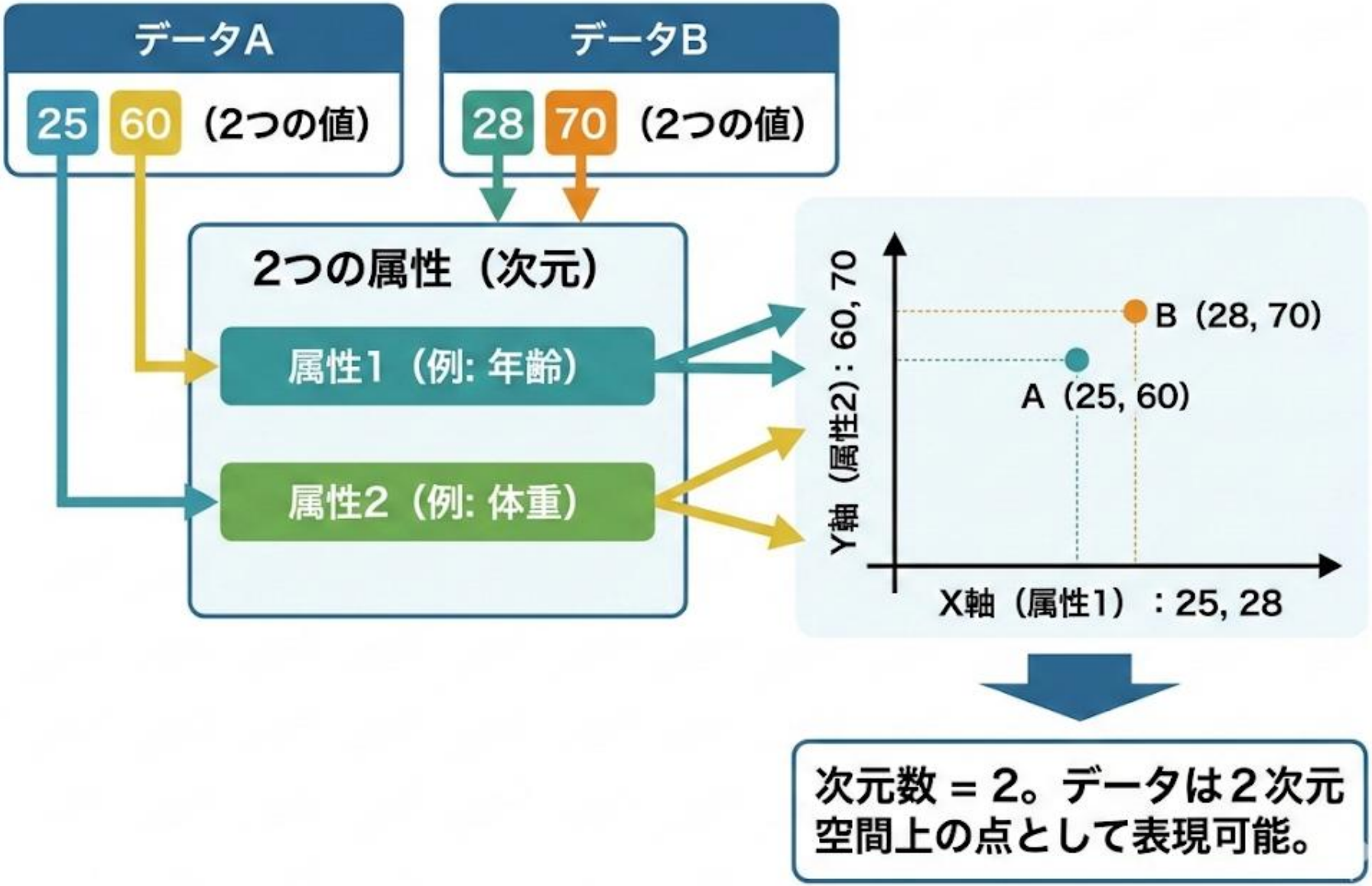
GOAL

データ分析の達人へ！

- 多次元データの可視化と分析
- データ品質の評価手法の習得
- スキルが飛躍的に向上

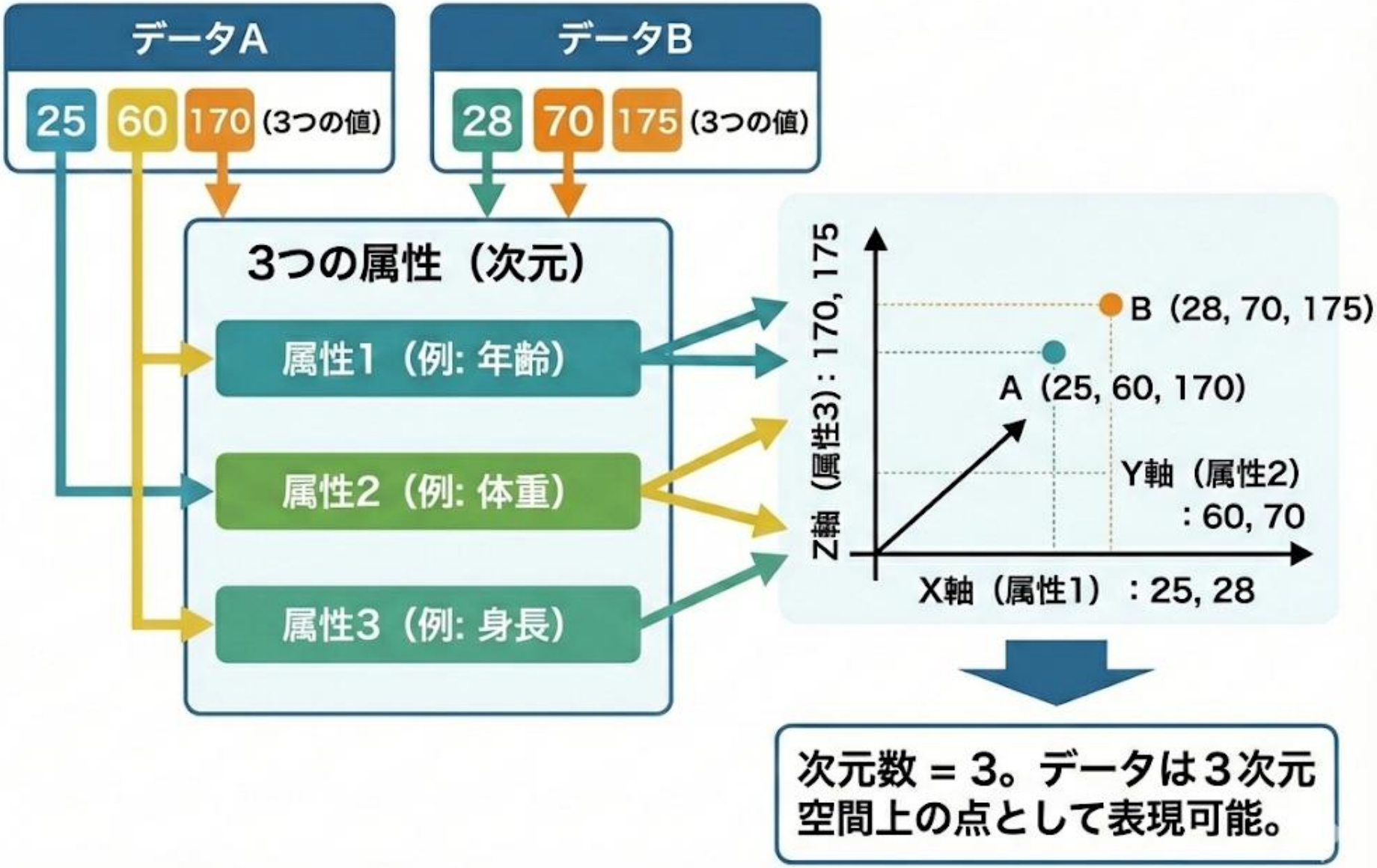
図解は Nano Banana Pro を用いて作成

データの次元数（2次元）



図解は Nano Banana Pro を用いて作成

データの次元数（3次元）

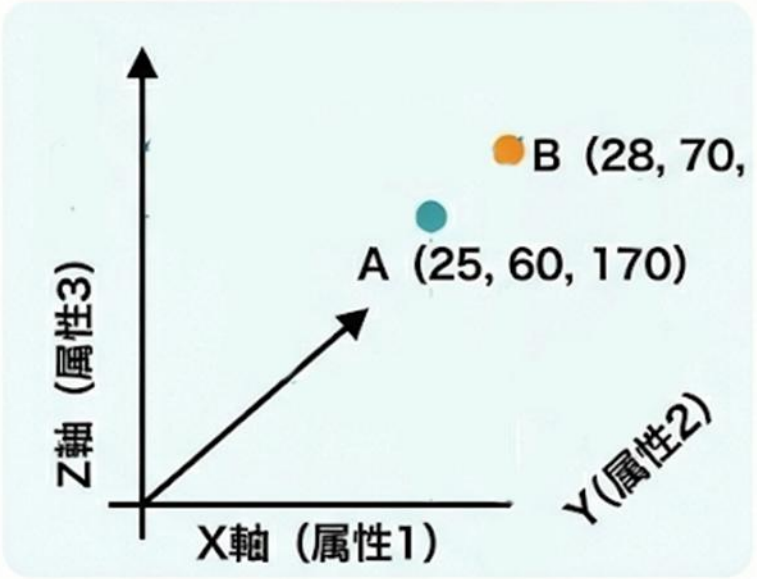


図解は Nano Banana Pro を用いて作成

データの次元削減（3次元から2次元）

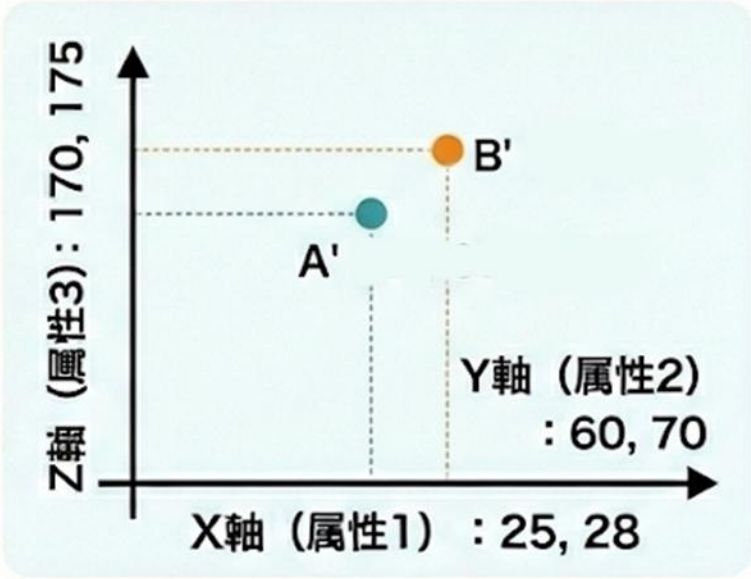
次元削減前（3次元）

25 60 170 28 70 175



次元削減
（例：属性3を削除）

次元削減後（2次元）

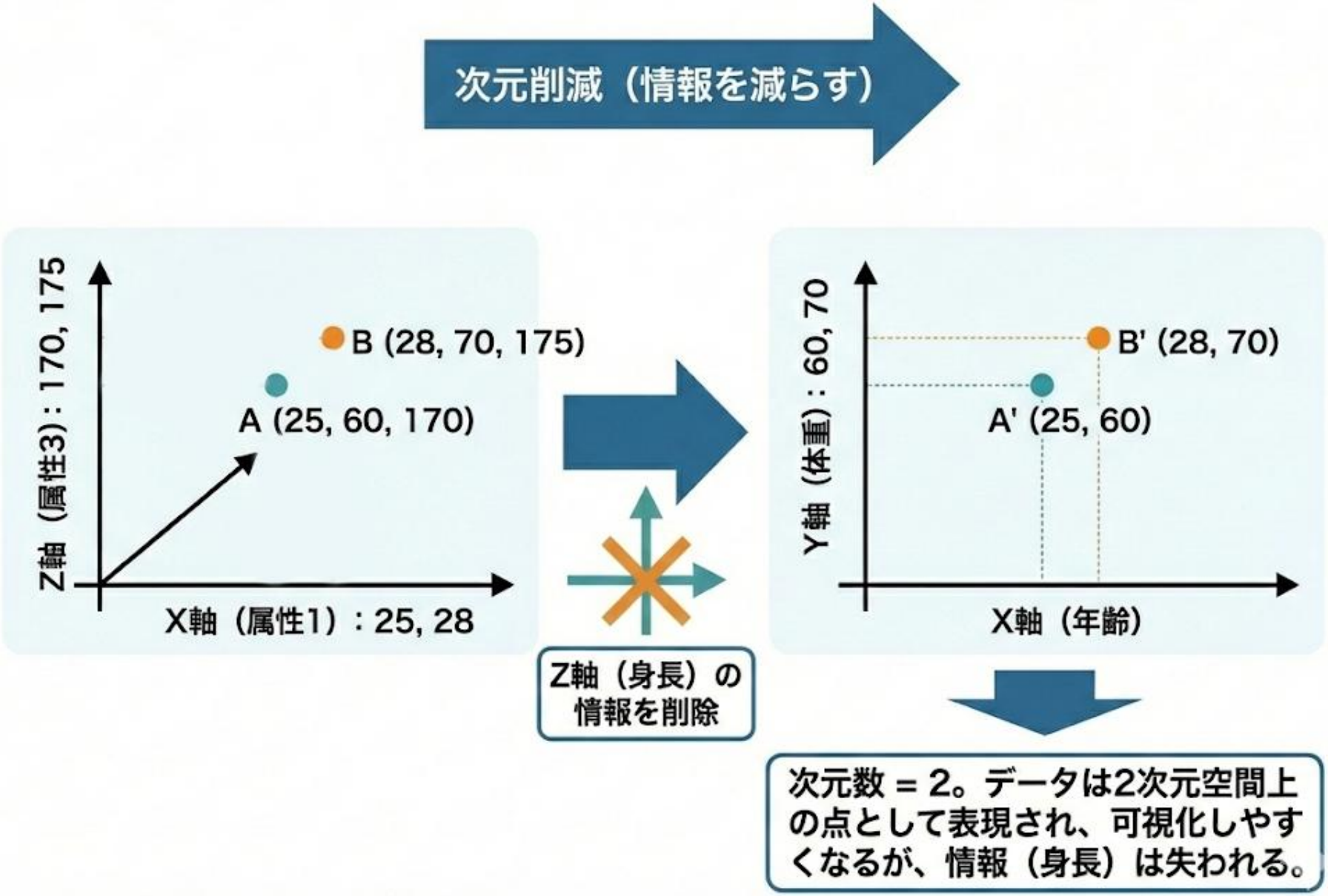


情報量：多、計算：複雑

情報量：少、計算：容易、
視覚化：簡単

図解は Nano Banana Pro を用いて作成

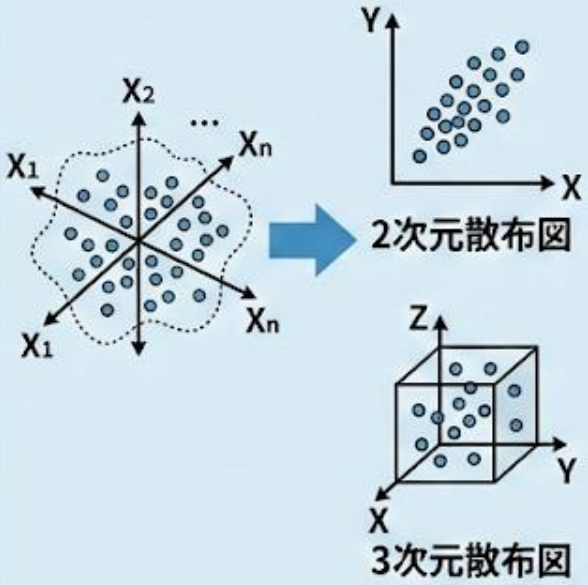
次元削減（3次元 → 2次元）



図解は Nano Banana Pro を用いて作成

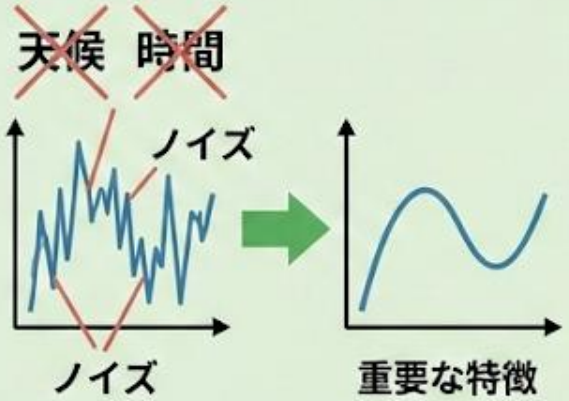
次元削減の効果

可視化



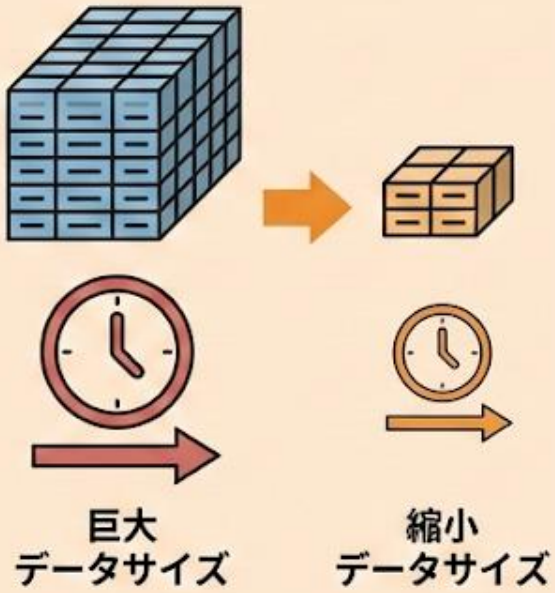
高次元を2D/3Dへ削減。
グラフ化可能に。

本質でない情報の除去



ノイズ・不要情報削除。
本質抽出。

計算の効率化



データサイズ縮小。
処理速度向上。

図解は Nano Banana Pro を用いて作成

次元削減の単純な方法

属性の削除

3次元データ

属性A	属性B	属性C

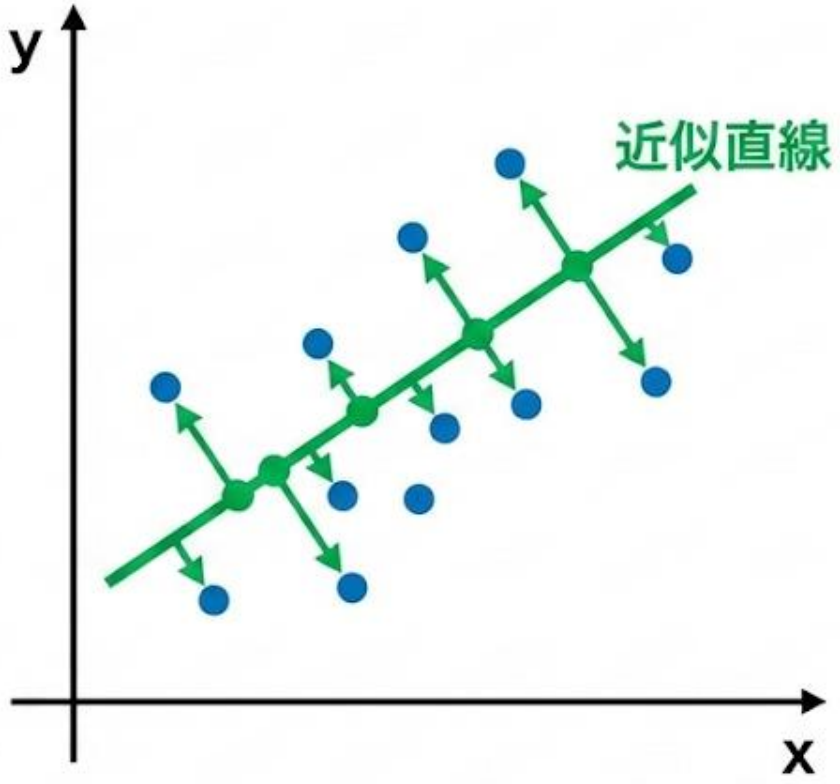
2次元化

属性A	属性B

不要な列を削除。

近似直線への投影

2次元データ

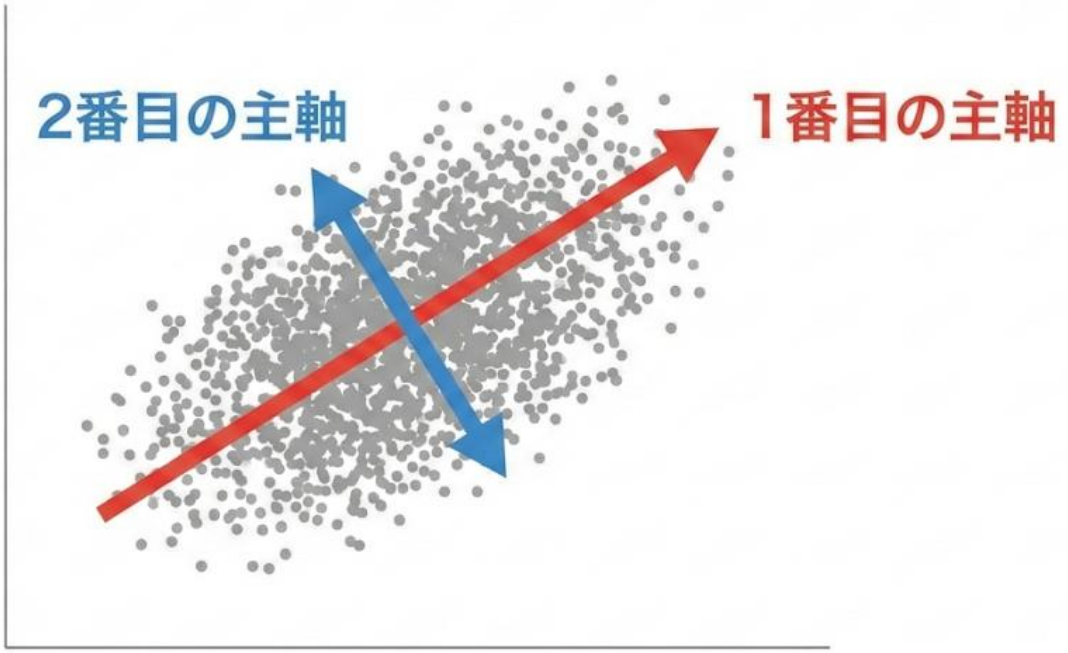


直線上に点を集約。

図解は Nano Banana Pro を用いて作成

主成分分析と主軸

主軸：主成分分析において、データの分散が最大となる方向の軸

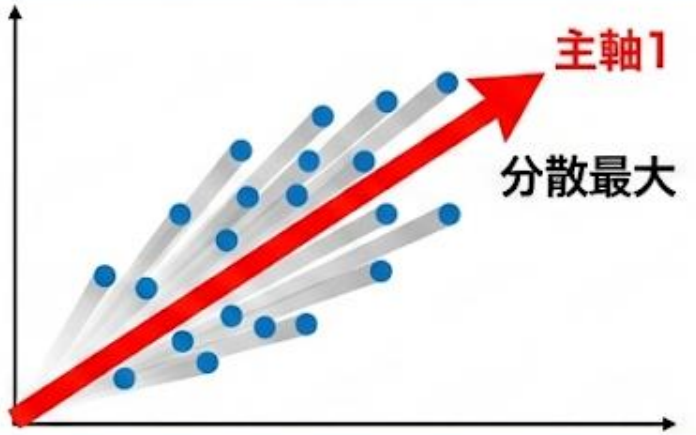


次元数は2
主軸を2つ

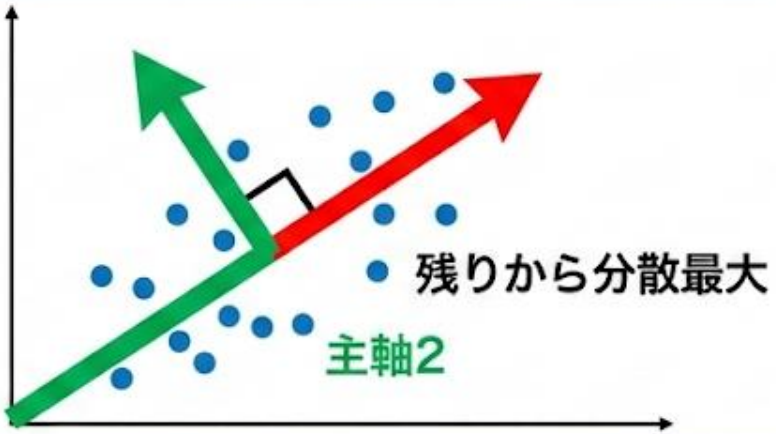
図解は Nano Banana Pro を用いて作成

主成分分析と主軸

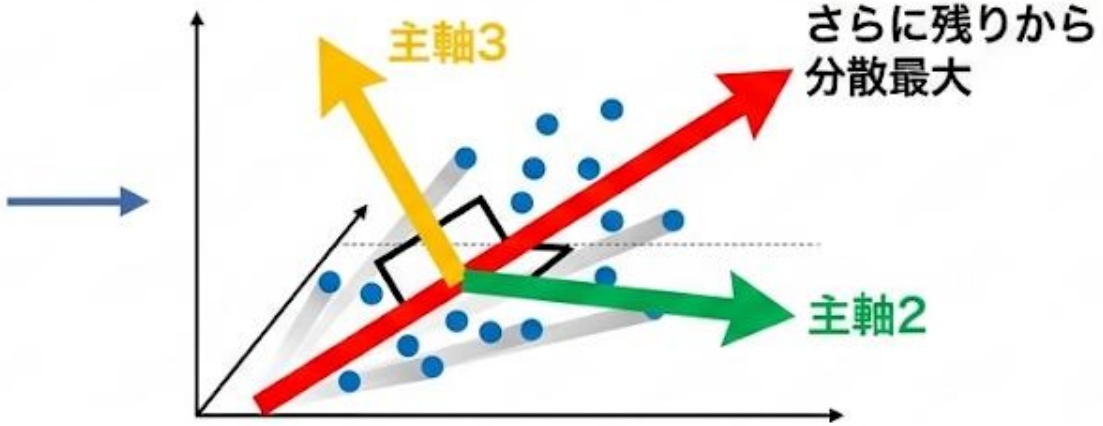
第1主軸：分散が最大となる方向



第2主軸：第1主軸を除去後、分散最大



第3主軸以降：同様に、残りから分散最大を選択



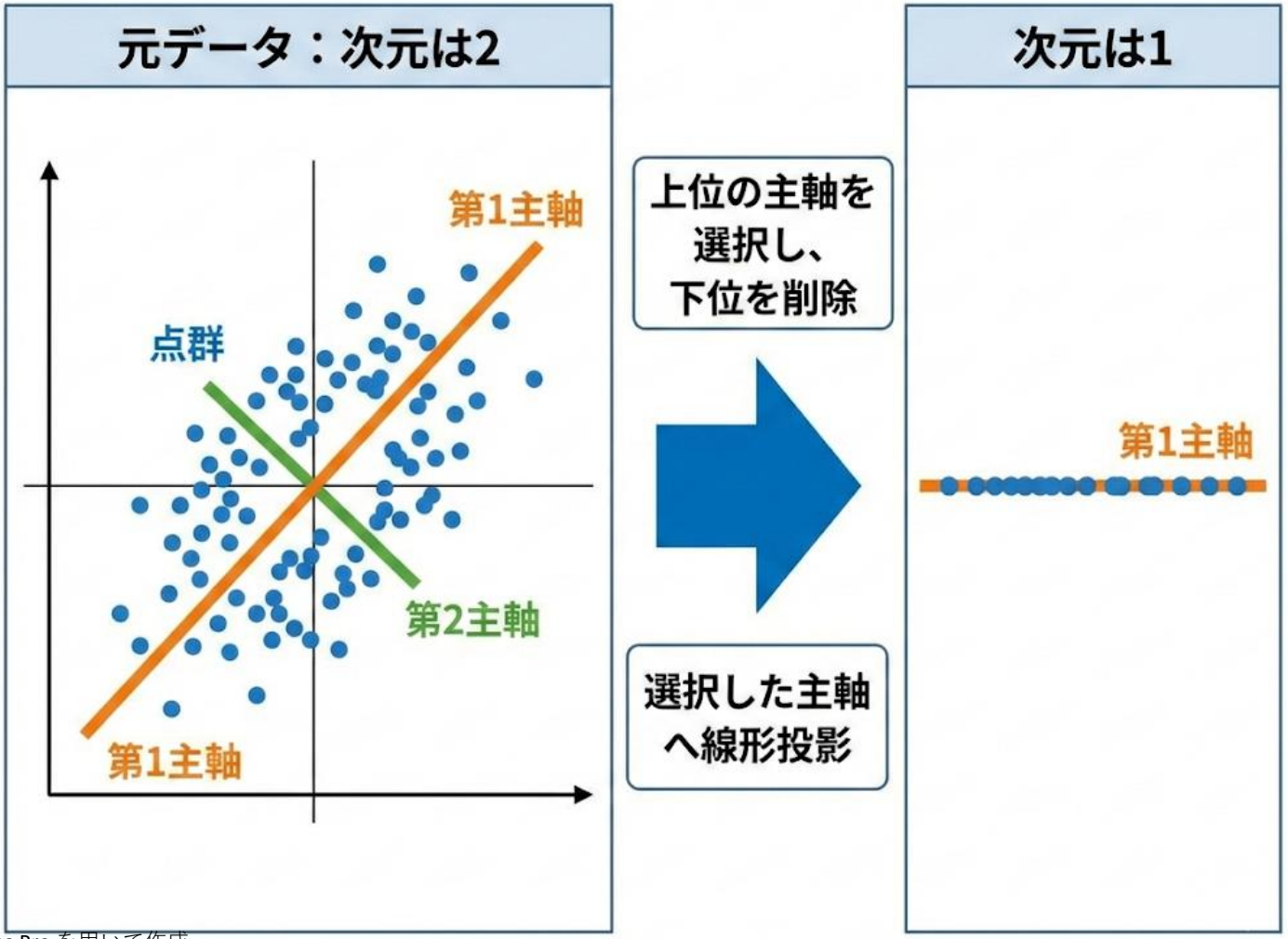
各主軸は互いに直交

※元のデータの次元数と
同じ数の主軸を作成可能

図解は Nano Banana Pro を用いて作成

主成分分析による次元削減

得られた主軸の中から上位の主軸を選択し、下位の主軸を削除



図解は Nano Banana Pro を用いて作成

主成分分析（PCA）の流れとPython実装

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA

# データ準備 (data, x, y)
pca = PCA(n_components=2)
pca.fit(data)

components = pca.components_

explained_variance = pca.explained_variance_

# 可視化
plt.axis('equal')
plt.scatter(x, y)

plt.quiver(..., color='r')

plt.quiver(..., color='b')

plt.show()
```

<- データから主成分（主軸）を学習

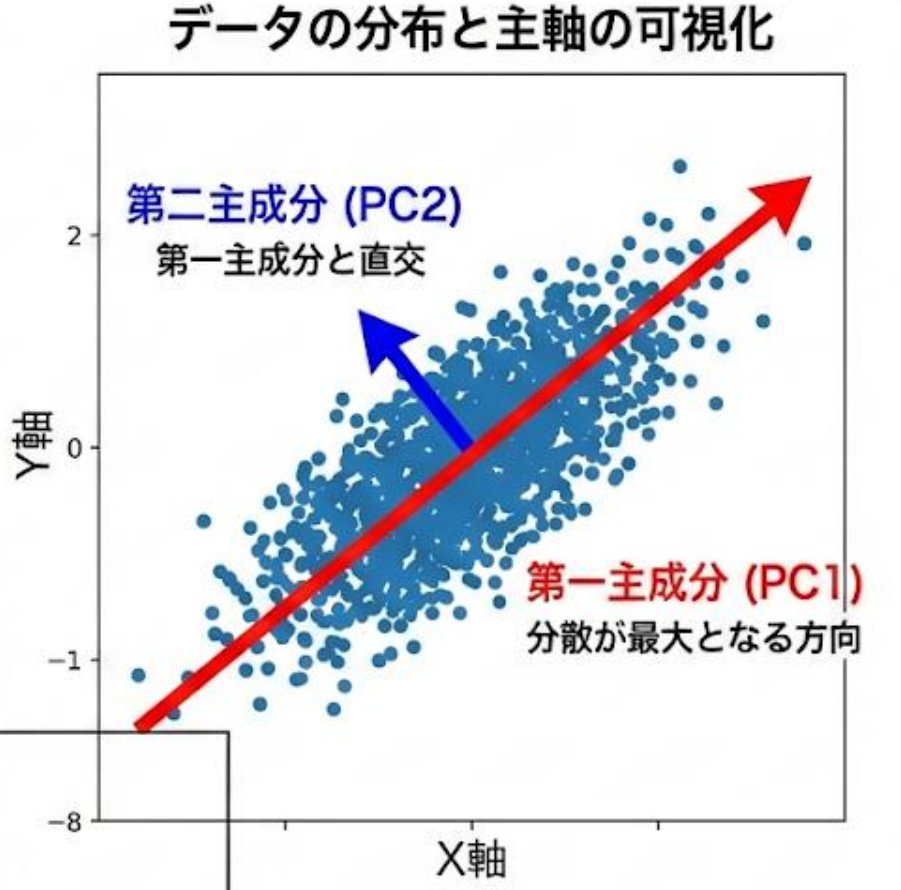
<- 主軸ベクトルを取得

<- 各主成分の分散を取得

<- 元データを散布図で表示

<- 第一主成分（赤矢印）

<- 第二主成分（青矢印）



次元削減：重要な情報（主成分）を残し、データを要約

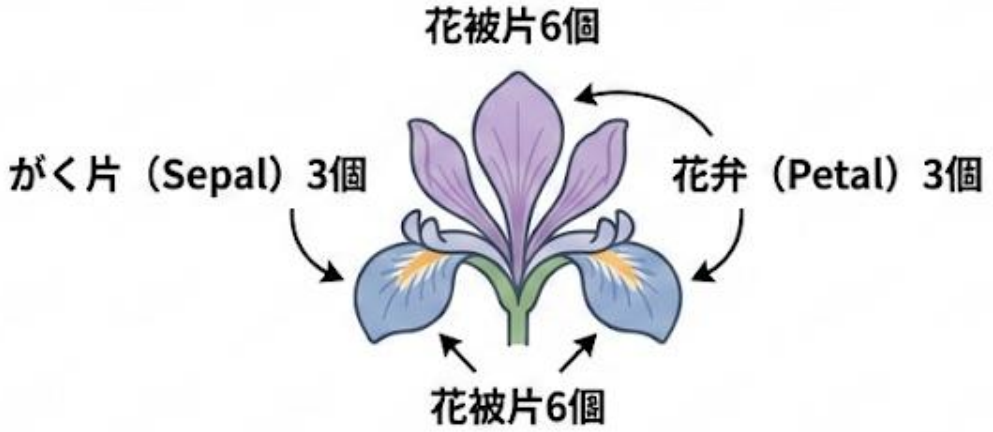
分散：データの広がり具合。大きいほど情報量が多いと解釈

図解は Nano Banana Pro を用いて作成

Irisデータセット概要

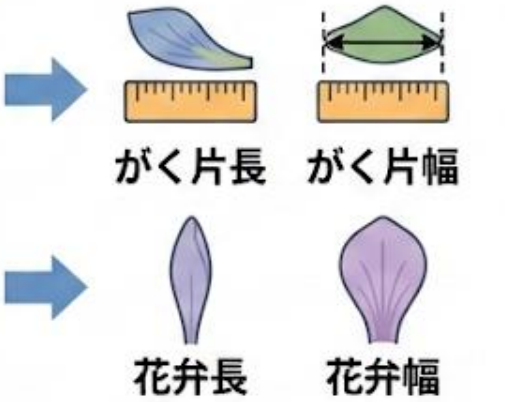
アヤメ属（Iris）

多年草



Irisデータセット

対象3種



がく片、花弁の幅・長さ計測

データ数：150
（50 × 3種）

1936年

作成者：Ronald Fisher

図解は Nano Banana Pro を用いて作成

Irisデータセット：特徴量と種類（ラベル）

`x`（特徴量データ：2次元配列）

がく片の長さ (cm)	がく片の幅 (cm)	花弁の長さ (cm)	花弁の幅 (cm)
5.1	3.5	1.4	0.2
4.9	3.0	1.4	0.2
6.4	3.2	4.5	1.5
5.1	3.0	1.6	0.3
4.7	3.0	4.6	1.5
⋮	⋮	⋮	⋮
6.4	3.2	4.5	1.5

各行＝1つのアヤメの測定値（4つの値）

`y`（ターゲットデータ：1次元配列）

0
0
1
2
0
1
2
⋮
...

0 = Setosa（種類A）

1 = Versicolor（種類B）

2 = Virginica（種類C）

各値＝xに対応する正解ラベル（アヤメの種類）

データセットの構成：特徴量（x）＋正解ラベル（y）

機械学習で分類を学習するためのペアデータ

図解は Nano Banana Pro を用いて作成

Irisデータセットの PCA：スケーリングの重要性と次元削減

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.datasets import load_iris
from sklearn.preprocessing import StandardScaler

# データ読み込み
iris = load_iris()
data = iris.data
y = iris.target

# スケーリング（重要！）
scaler = StandardScaler()
data_scaled = scaler.fit_transform(data)

# PCAによる次元削減
pca = PCA(n_components=2)
data_pca = pca.fit_transform(data_scaled)

# 可視化
plt.scatter(data_pca[:, 0], data_pca[:, 1], c=y)
plt.title('PCA of IRIS dataset')
plt.show()
```

スケーリング：異なる尺度の特徴を揃える。分散算出に必須

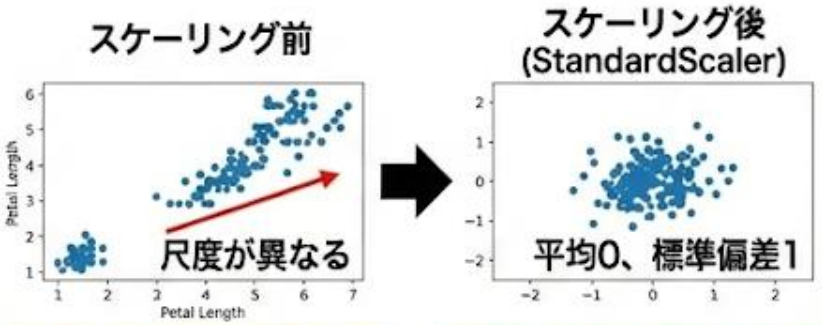
<- 平均0, 標準偏差1へ

<- 各特徴量の尺度を統一

<- 2次元データへ射影

<- クラス別に色分けして表示

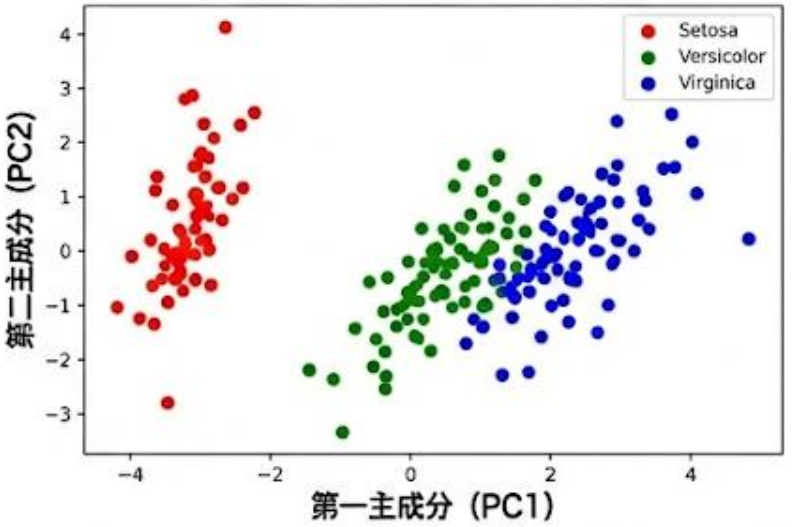
スケーリングの重要性（概念図）



分散が偏り、主成分が正しく求まらない可能性

各特徴量が公平に扱われ、適切な主成分が得られる

PCA後の可視化（2次元）

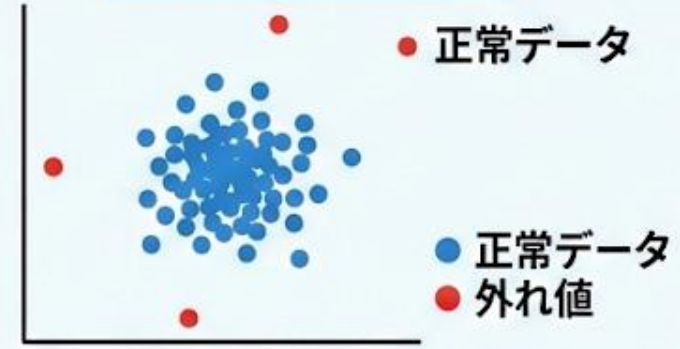


結果：スケーリングを経て、Irisデータセットの構造（クラスの分離）が2次元で可視化された

図解は Nano Banana Pro を用いて作成

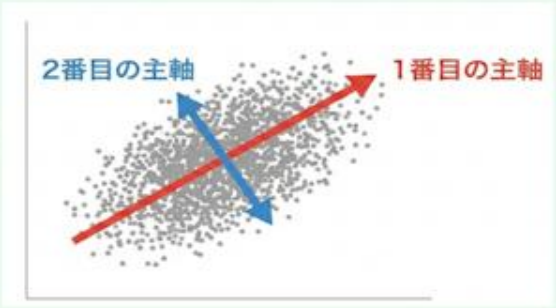
主成分分析（PCA）を用いた外れ値検出の手順

1. 元データ入力



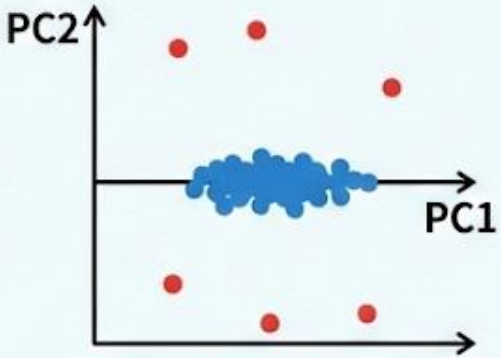
多次元データ（外れ値含む）

2. 主成分分析（PCA）実行



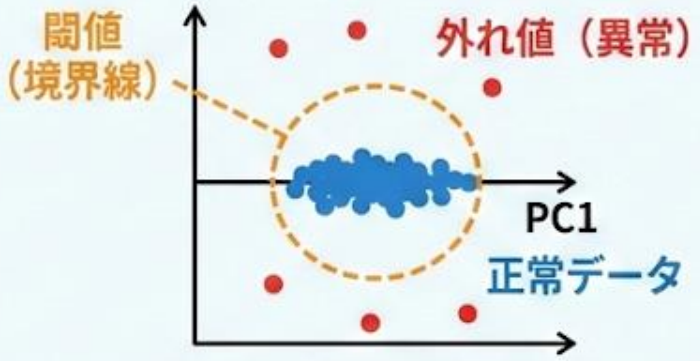
次元圧縮・特徴抽出

3. 処理結果の確認



外れ値が分離・可視化

4. 外れ値の区別・判定

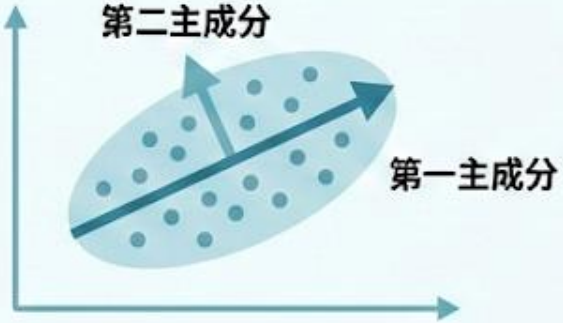


閾値に基づく判定

図解は Nano Banana Pro を用いて作成

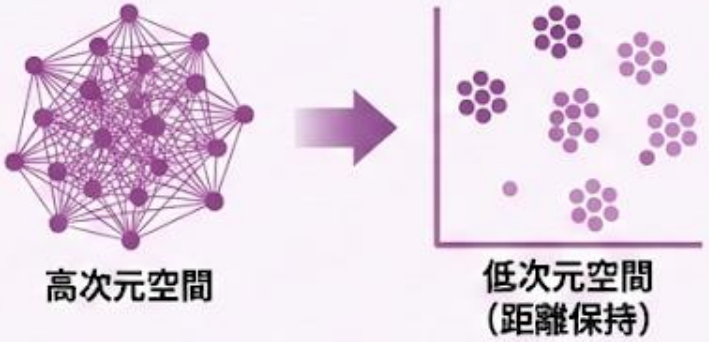
次元削減のさまざまなアプローチ

主成分分析（PCA）



分散最大化、特徴軸の発見。
データの広がりを最も表す軸を抽出。

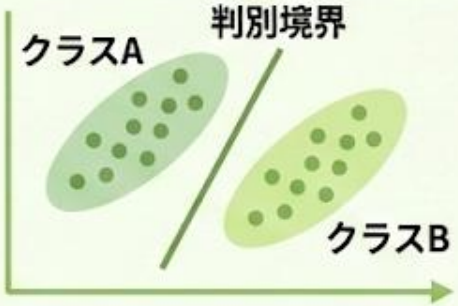
t-SNE



データ間距離の保持。
高次元での近さを低次元でも維持。

線形判別分析（LDA）

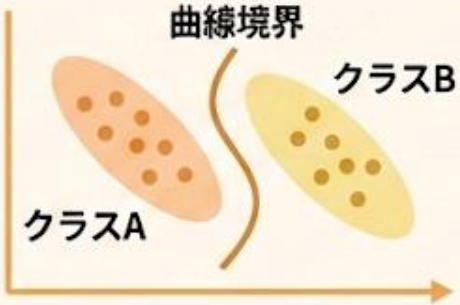
教師あり



教師あり、クラス分離。
各クラスが正規分布、共通の共分散を仮定。

改良

二次判別分析（QDA）



LDAの改良版。
共通の共分散を仮定しない。境界が曲線に。

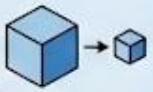
図解は Nano Banana Pro を用いて作成

主成分分析と次元削減

基礎概念

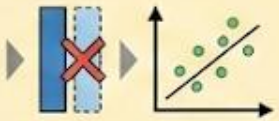
データの次元
必要な情報の数

次元削減
次元を減らす処理

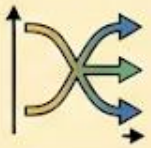


次元削減の手法

単純な方法
属性削除、投影



高度な手法
主成分分析など

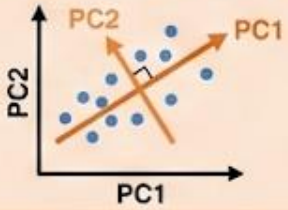


次元削減の効果

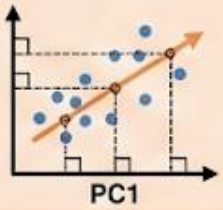
- 可視化
- 情報除去
- 計算効率化

主成分分析の原理

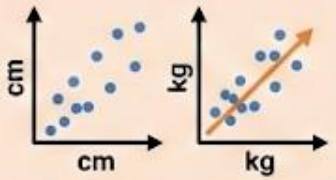
主軸決定
分散最大・直交



投影
上位主軸へ

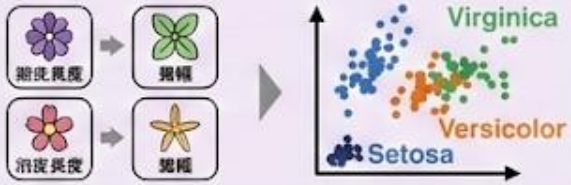


スケーリング
尺度統一が重要

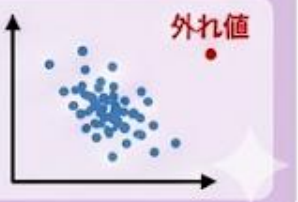


実践と応用

Irisデータセット
4次元 → 2次元



外れ値検出
応用：外れ値と
主成分分析



図解は Nano Banana Pro を用いて作成