ae-4.機械学習の基礎②

ー　教師なし学習編　ー

（AI演習）（全１５回）

https://www.kkaneko.jp/ai/ae/index.html

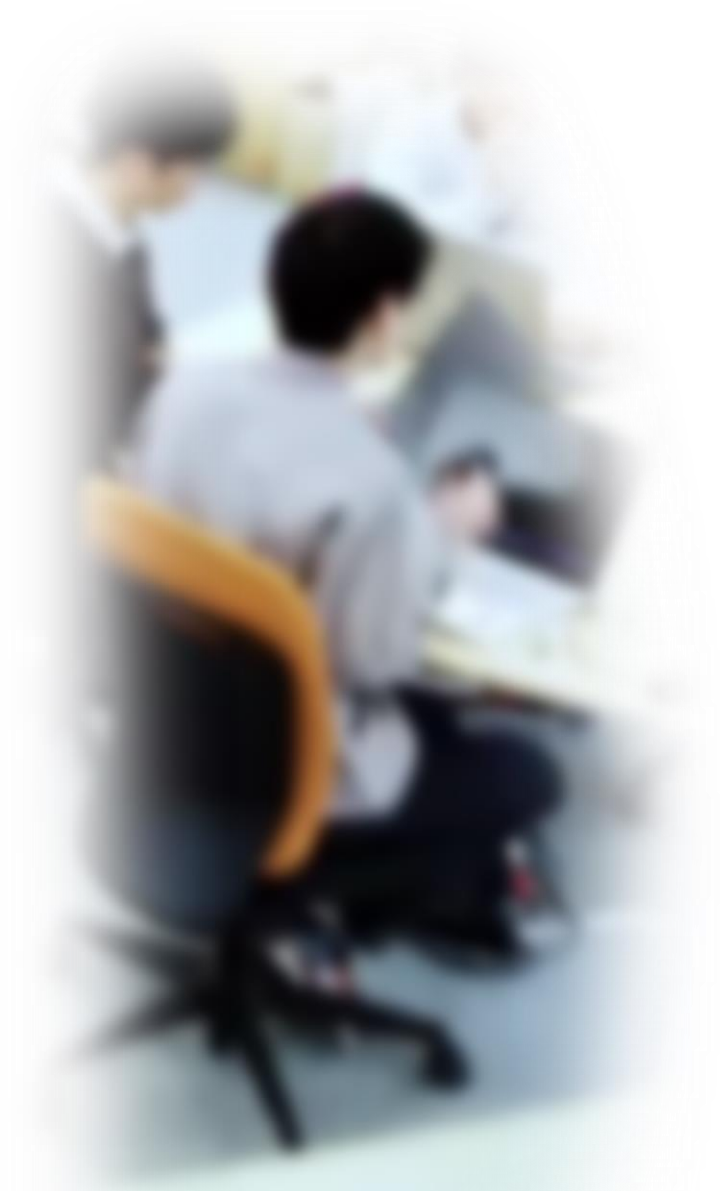
金子邦彦

人工知能（AI）の学習のための指針

1. **実践重視**： AIツールを実際に使用し、機能に慣れる

2. **エラーを恐れない**： 実行においては、エラーの発生の可能性がある．エラーを恐れず，むしろ学習の一部として捉えるポジティブさが大切．

3. **段階的学習**：基礎から応用へと段階的に学習を進め，AIの可能性を前向きに捉える



アウトライン

1. イントロダクション
2. 教師無し学習
3. クラスタリング
4. オートエンコーダ
5. 画像生成AIとの関連
6. 学習のパリエーション

Python

- **Python** は多くの人々に利用されている**プログラミング言語**の１つ
- **読みやすさ，書きやすさ，幅広い応用範囲**が特徴

```
from keras.models import Sequentia
model = Sequential()
from keras.layers import Dense, Ac

model.add(Dense(units=64, input_di
model.add(Activation('relu'))
model.add(Dense(units=max(set(y_tr
model.add(Activation('softmax'))
model.compile(loss='sparse_categor
              optimizer='sgd',
              metrics=['accuracy']
model.fit(x_train, y_train, epochs
score=model.evaluate(x_test, y_tes
print(score)
model.predict(x_test)
model.summary()
```

```
poch 1/200
'3 [==============================] - 0s
 3200
 och 2/200
 3 [==============================] - 0s
 200
  ch 3/200
   [==============================] - 0s
  0
```

Python 言語が広く使用されている理由

文法のシンプルさ

- **Python** は，**直感的で読みやすい文法**
- 例えば、**print** で簡単に**出力**できる、**if** や **else** で**条件分岐**、**for** や **while** で**繰り返し（ループ）**

拡張性

- **多岐にわたる分野で利用が可能**
- 例えば、**関数やクラスを定義**するための **def** や **class**、**継承**や**オブジェクトの属性名と値**を操作するための **super** や **vars** などがある。

柔軟性

- **シンプルなスクリプトも、高度なプログラムも作成可能**
- **オブジェクト指向**の機能を持ち、**__init__** や **self** のようなキーワードを使用して**クラス**を利用できる。

trinket

- **Trinket** は**オンライン**の **Python、HTML** 等の**学習サイト**
- ブラウザで動作
- 有料の機能と無料の機能がある
- **自分が作成した Python プログラムを公開し、他の人に実行してもらうことが可能**（そのとき、書き替えて実行も可能）

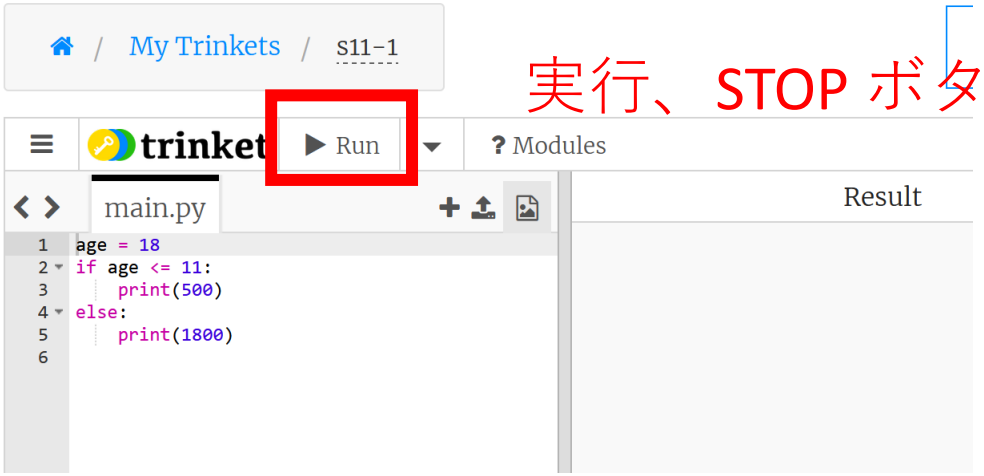
- **Python の標準機能**を登載、その他、次の外部ライブラリがインストール済み

matplotlib.pyplot, numpy, processing, pygal

trinket でのプログラム実行

- **trinket は Python, HTML などのプログラムを書き実行できる**サイト
- https://trinket.io/python/0fd59392c8

のように、違うプログラムには違う URL が割り当てられる



- 実行が開始しないときは、「**実行ボタン**」で**実行**
- ソースコードを**書き替えて再度実行**することも可能

4-1. イントロダクション

機械学習の基本

機械学習は，**コンピュータ**が**データ**を使用して**学習**することにより**知的能力を向上**させる技術

- **情報の抽出**：データからパターンや関係性を自動で見つけ出す能力を持つ
- **知的なタスクの実行**：予測，分類などの知的なタスクを実行

機械学習の３つの特徴

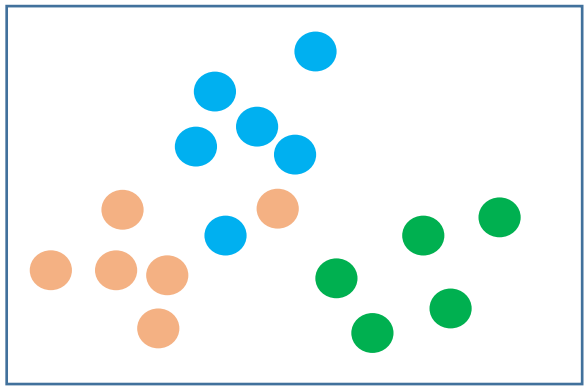
1. **情報の抽出**：データからパターンや関係性を自動で見つけ出す能力を持つ
2. **簡潔さ**：人間が設定していたルール等を，自動生成できる
3. **限界の超越**：従来の方法では困難だった課題も解決できる可能性がある

機械学習における訓練データの役割

- **訓練データ**は，**機械学習の学習に使用されるデータ**である．
- **データを用いた学習**により，コンピュータは分類や予測などの**知的な能力を獲得**する．

  例：画像分類では分類済みの大量データを使用する．

訓練データ



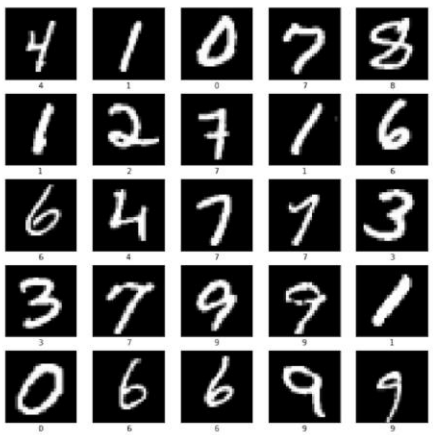
3種類に分類済み

学習

学習者

機械学習の学習プロセス

①データの準備：
目的に応じた訓練データを準備



画像 60000枚
（うち一部）

4 1 0 7 8
1 2 7 1 6
6 4 7 7 3
3 7 9 9 1
0 6 6 9 9

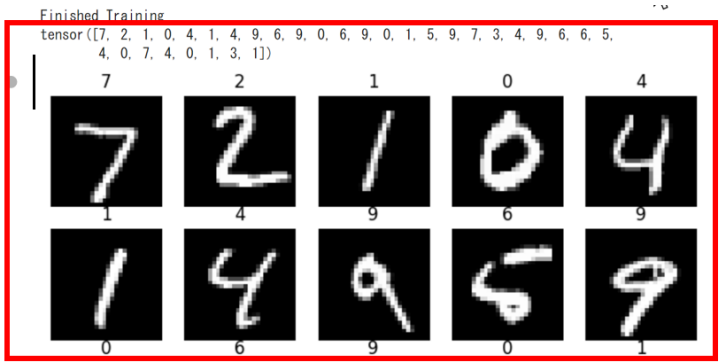
正解 60000個
（うち一部）

②学習の実行：
データを用いてパターンを学習.
モデルを構築

プログラム

データを用いて学習を行う．学習の結果、文字認識の能力を獲得.

③タスクの実行：
新しいデータを処理



機械学習まとめ

機械学習の特徴

- データを用いて知的能力を向上
- 自動でデータのパターンを抽出
- さまざまなタスクを自動実行

応用事例

画像理解、自然言語処理、予測など多数

機械学習の定義：

- **訓練データ**を用いて**学習**し、その結果として知的能力が向上
- **訓練データの追加**により、さらに**知的能力が向上**する可能性

機械学習の特徴と応用

- **能力の多様性**:

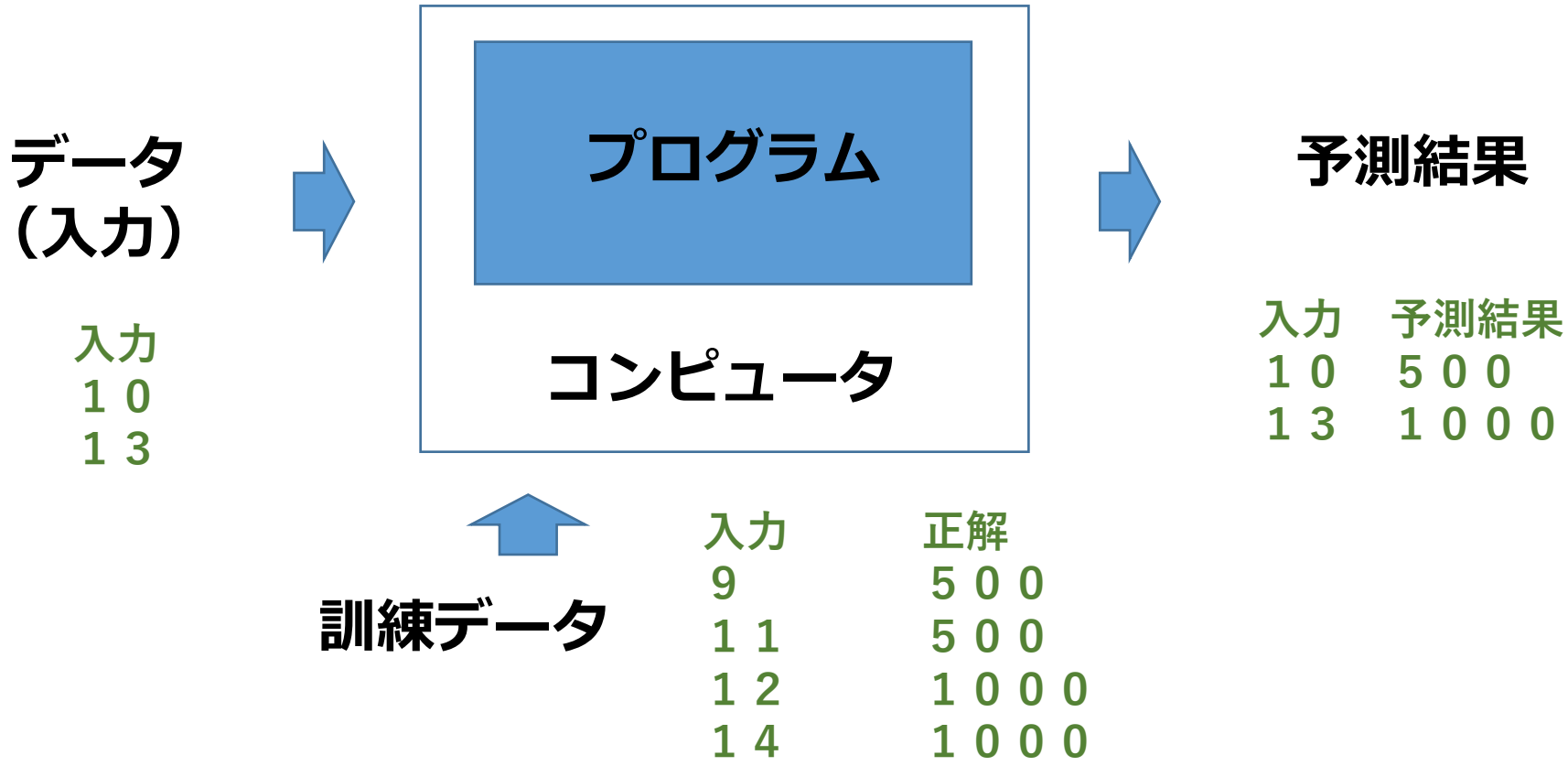
分類、検出、識別、認識（文字認識、画像の認識、音声認識）、予測、合成、翻訳、特徴抽出、ランク付け、情報推薦など、**さまざまな能力を持つ**

- **応用範囲の広さ**:

経済、金融、ロボット、医療、医学研究（遺伝子解析）など、様々な分野で問題解決に活用

機械学習は多様な能力を持ち、経済や金融から医療やロボットまで広範な応用が可能

機械学習によるプロセス



- 機械学習では，訓練データから自動的にパターンや関連性を学習する．
- 学習結果のモデルを利用して、新しいデータ（未知のデータ）に対して，タスクが実行される．人手による規則の記述が不要である．

汎化の概念と特徴

- **汎化**は，**訓練データを基に学習**し，**未知のデータも適切に処理**する能力である.
- 機械学習の重要な特徴.
- 汎化の結果は必ずしも正解とは限らない.

汎化

訓練データ

入力	正解
9	500
11	500
12	1000
14	1000

訓練データと未知のデータ

入力	予測結果
7	**500**
8	**500**
9	500
10	**500**
11	500
12	1000
13	**1000**
14	1000
15	**1000**
16	**1000**

汎化は，**訓練データを基に学習**し，
未知のデータも適切に処理する能力

まとめ

- **機械学習の基本**：データを用いた学習による知的能力の向上
- **訓練データの重要性**：学習は訓練データに依存する
- **汎化**：**未知データへの対応**
- 実践的応用：様々な分野での活用可能性

4-2. 教師なし学習

教師なし学習

- **データのパターン，構造，隠れた特徴や関係性を見つける**
- 教師あり学習と異なる仕組みを持つ.

  教師あり学習：**ラベル付きデータ**を使用

  教師なし学習：**ラベルのないデータ**を使用
- ラベル（正解、目標を表す）なしに学習を行う
- **人間が気づいていない複雑なパターンを発見可能である**

  例：顧客の購買パターン分析、異常検知システム

教師あり学習と教師無し学習の比較

	教師あり学習	**教師なし学習**
訓練データ	ラベル付きデータ	ラベルなしデータ
用途	予測・分類	パターン発見
特徴	明確な正解がある	正解が不明確

- **教師あり学習**：入力データと正解（ラベル）のペアを使用
  例：有害メール分類
- **教師無し学習**：ラベルのないデータのみを使用
  例：顧客セグメンテーション）

教師なし学習はより柔軟で予想外の洞察を得られる可能性がある

教師なし学習の主要タスク

- **クラスタリング（似た特徴を持つデータをグループ化）：**

  例：購買行動に基づく顧客グループ分け

- **異常検知：**

  例：クレジットカード不正利用の検出

- **推薦システム：**

  例：映画配信サービスの映画推薦

- **次元削減、データ圧縮：**

  例：高解像度画像の圧縮

教師なし学習のまとめ

- **教師なし学習の特徴**
  - ラベルのないデータからパターンや構造を発見する
  - 人間が気づいていない複雑な関係性を見出す可能性がある
- **教師あり学習との本質的な違い**
  - 教師あり：ラベル付きデータで予測・分類
  - 教師なし：ラベルなしデータでパターン発見
- **主要な応用分野**
  - クラスタリング
  - 異常検知
  - 推薦システム
  - データ圧縮

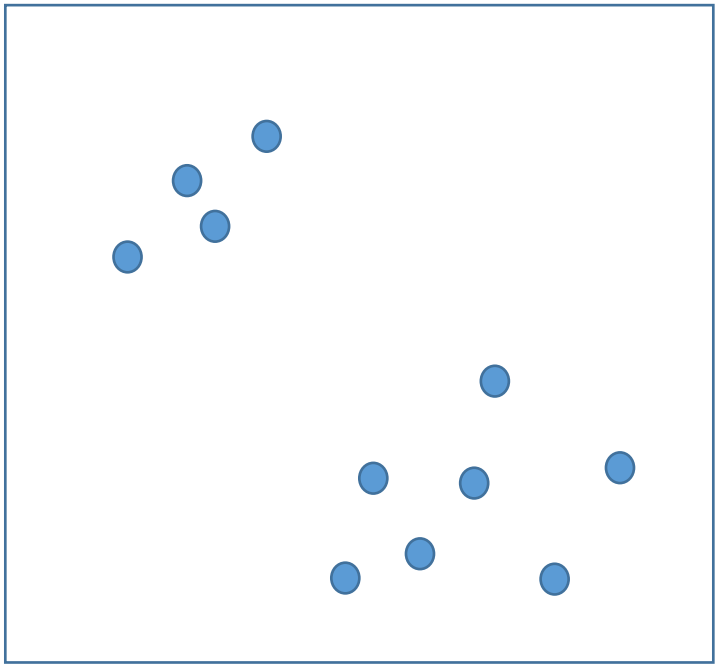
4-3. クラスタリング

クラスタリングの基本概念

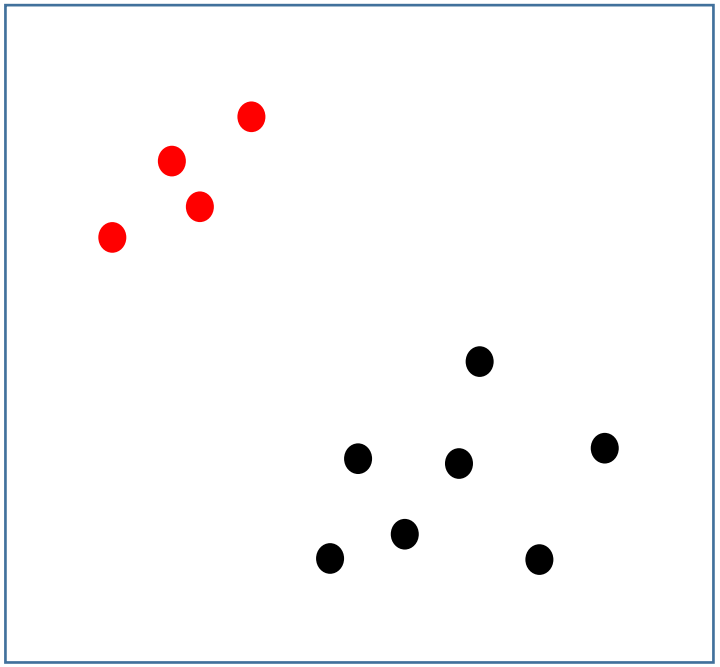
- **クラスタ内のデータポイントは互いに近い**
- **異なるクラスタのデータポイントは離れている**
- クラスタの個数はさまざま（2, 3, 4, 5, …）に設定可能

クラスタリングは教師なし学習の一種である

データ



２つのクラスタ



クラスタリングの基礎

- **データポイント**を**グループ（クラスタ）に分類**
- **似たデータポイント**を**同じクラスタ**に
- **異なるデータポイント**を**別のクラスタ**に
- **データ**の**自然な構造を理解する**のに役立つ

  例：顧客データ分析

  例：マーケティング戦略立案

  例：Eコマースサイトでの商品カテゴリー自動分類

演習１．データの散布図

ページ２３～２６

- モール顧客データを**散布図**にする
  - **１つめ**：年齢と年間収入の**関係**
  - **２つめ**：年間収入と支出スコアの**関係**
- Matplotlib ライブラリを使用
- **全体の分布**を一目で把握できる

モール顧客データ（Mall Customers Data）

AI 学習に適する架空のデータセット２００名分

データ内容

- **顧客ID**
- **年齢**
- **性別**
- **年間収入**
- **支出スコア（行動基準）**

支出スコア（行動基準）
・0 から 100 の値
・各人の支出しやすさを示す

主なタスク

- クラスタリング実践
- 顧客の状況の把握

応用分野

- マーケティング戦略立案
- 顧客行動分析
- 市場分析

主な利用者

- データサイエンス・マーケティング分析の学習者

支出スコア

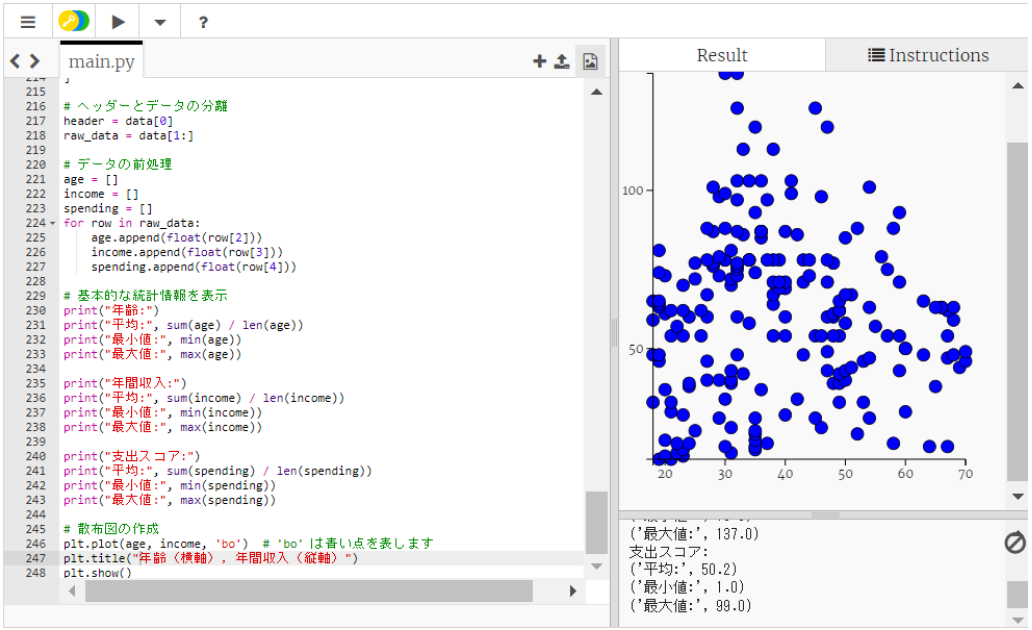
- 支出スコアの段階
  - **低いスコア（例：1-20）**：消費が少ない
  - **中程度のスコア（例：40-60）**：平均的な消費をする
  - **高いスコア（例：80-100）**：消費が多い
- 支出スコアを算出する手段
  - **購買頻度，購買金額，購入商品の種類，ポイントプレゼント等への参加度，特別セール等の行事への反応度**
- データ分析における意義
  - **行動予測**：将来の行動を予測．他のデータ（年間収入など）と組み合わせてより高い精度を狙う
  - **マーケティング戦略**：高スコアの顧客向けの特別オファー、低スコアの顧客のキャンペーンなど

年齢と年間収入

① trinket の次のページを開く

https://trinket.io/python/1445696bc0e0

② 実行結果が，次のように表示されることを確認



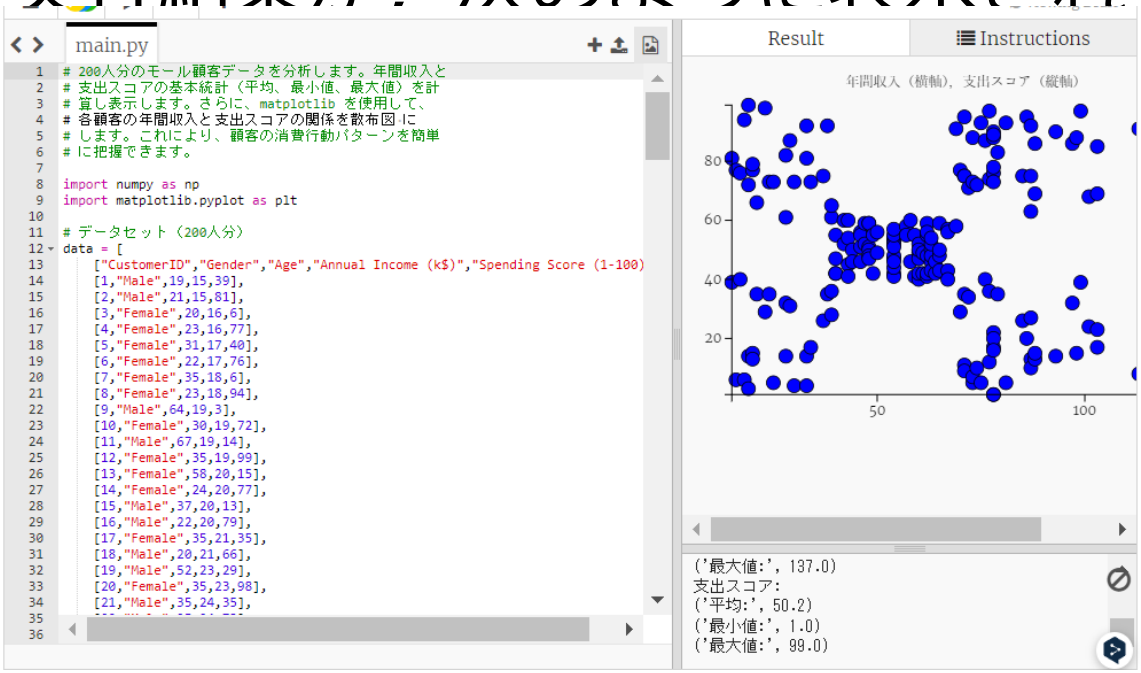
- 実行が開始しないときは、「**実行ボタン**」で**実行**
- ソースコードを**書き替えて再度実行**することも可能

年間収入と支出スコア

③ trinket の次のページを開く

https://trinket.io/python/afe760328290

④ 実行結果が，次のように表示されることを確認



- 実行が開始しないときは、「**実行ボタン**」で**実行**
- ソースコードを**書き替えて再度実行**することも可能

演習２．クラスタリング

ページ３０～３３

- モール顧客データの年間収入と支出スコアに対して，**クラスタリングを行う**

● クラスタ数を２，３，４，５，６と変化させて観察しよう！

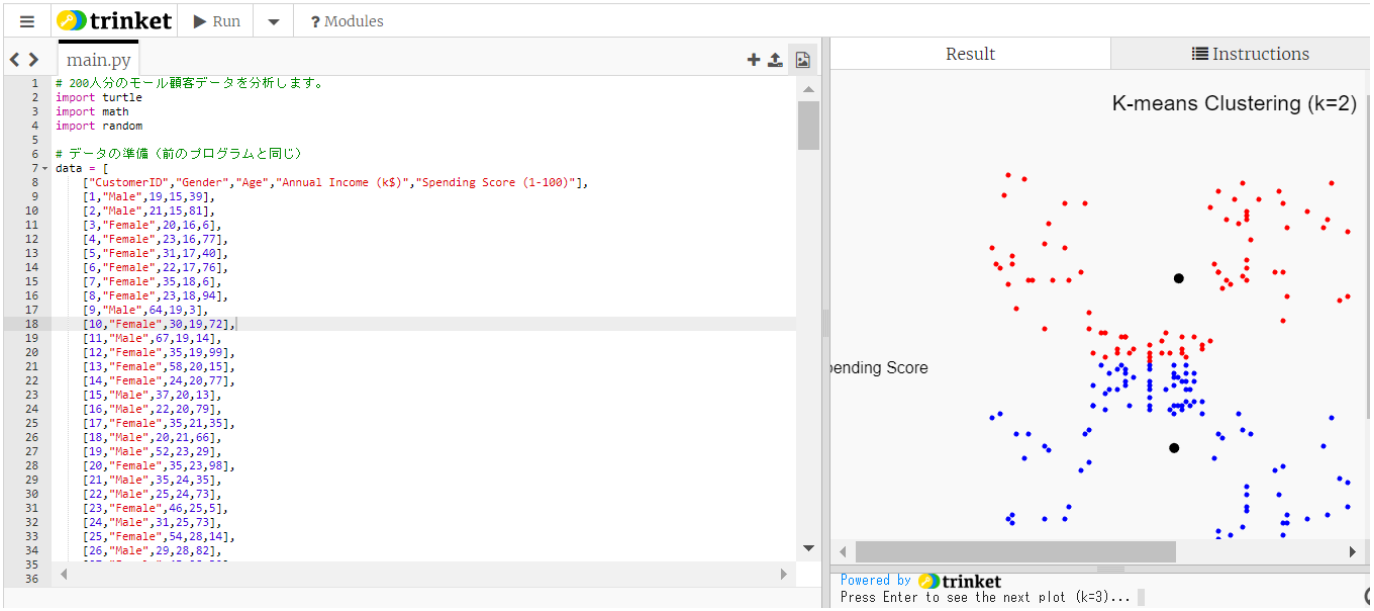
- さらに学びたい人は，自主的に演習３へ

クラスタリング

① trinket の次のページを開く

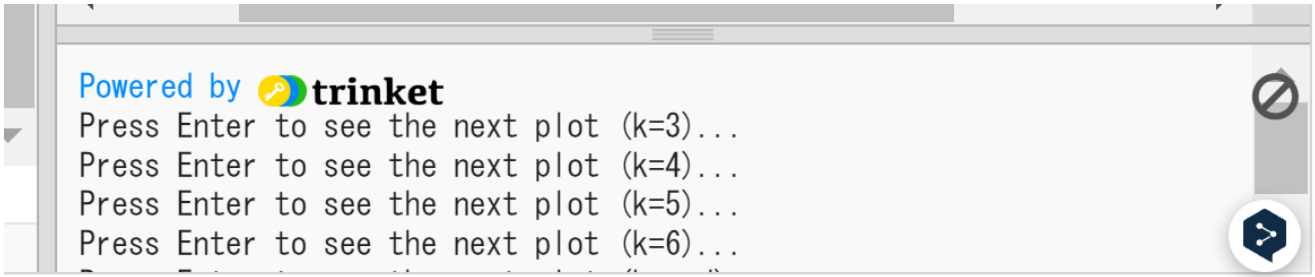
https://trinket.io/python/93ab2fb15488

② 実行結果が，次のように表示されることを確認



- 実行が開始しないときは、「**実行ボタン**」で**実行**
- ソースコードを**書き替えて再度実行**することも可能

③　右下をクリック．Enter キー



④　**クラスタ数**が 3 に**増える**ので確認

その後，③の操作と確認を繰り返す

演習３．クラスタリングをビジュアルに学べる学習サイト

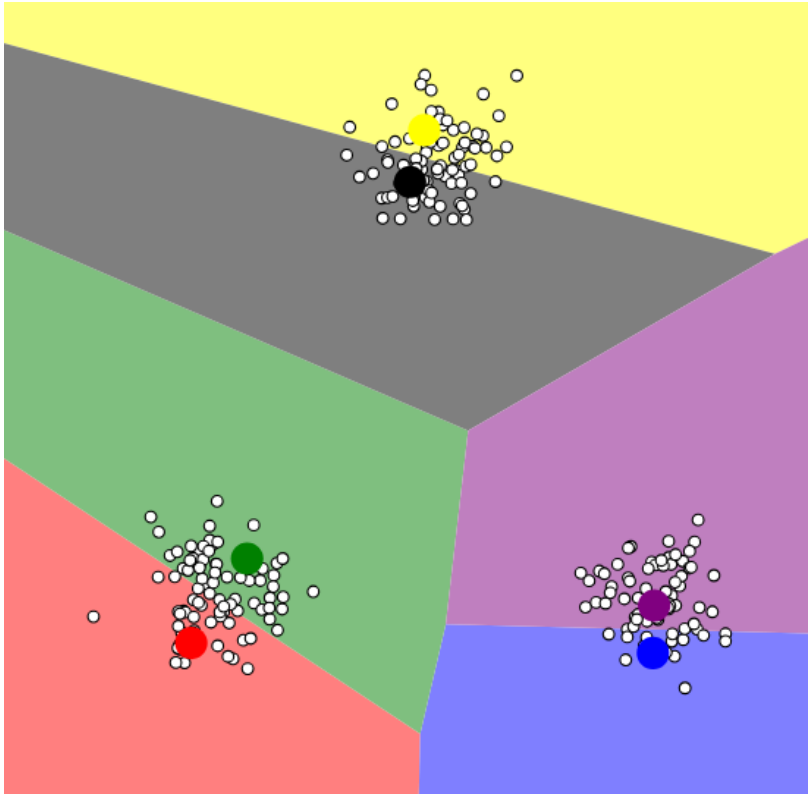
ページ３４～３５

クラスタリング学習サイトの活用

- サイトにアクセス:

https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

- **データセット**: 「**Randomly**」を選択
- **分布タイプ**: 好みのものを選択（例：**Gaussian Mixture**）
- 「**Add Centroid**」を**クリック**し、**結果を確認**。収束を観察しながら繰り返す
- **やり直す**場合は「**Restart**」をクリック



クラスタリングで用いた K-means 法

- 主なクラスタリング法

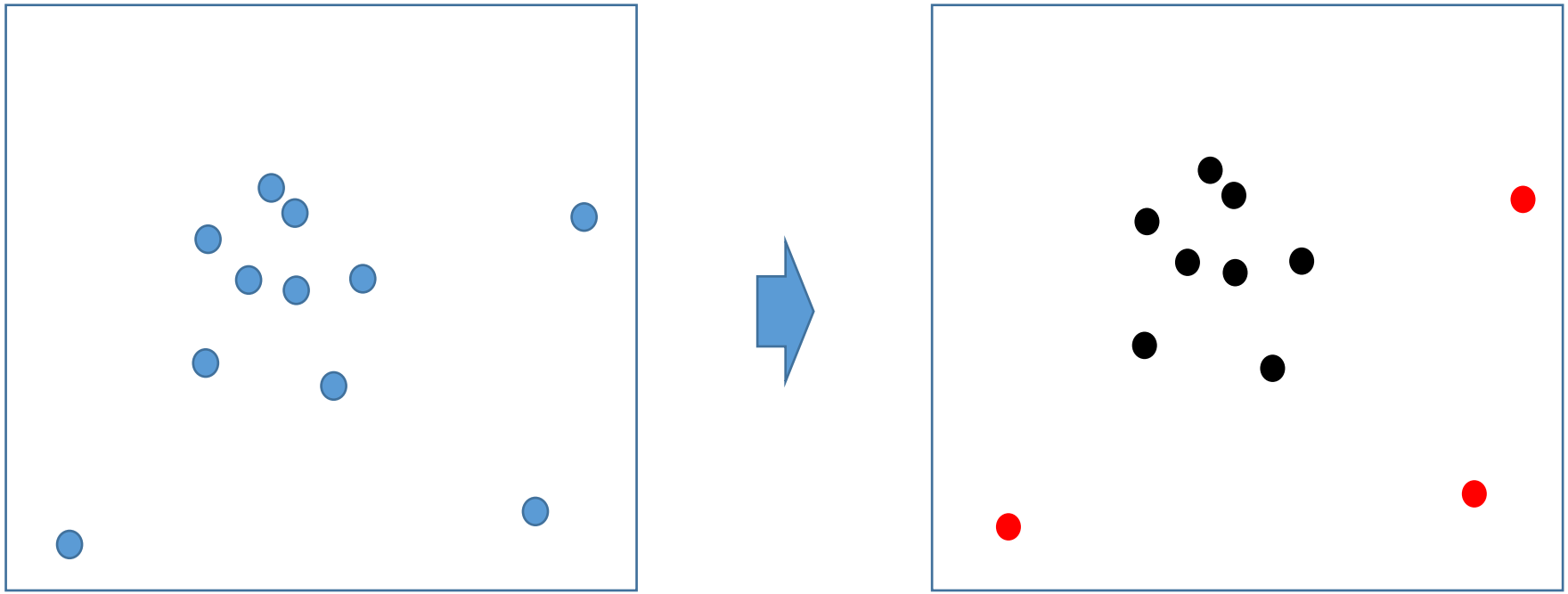
K-means （中心ベース），階層的クラスタリング，DBSCAN（密度ベース）など

- K-mean 法では，**中心点をランダムに初期化**し，その後，**各データポイントを，最も近い中心点に割り当て**

クラスタリングの異常検知への応用

正常なデータと異常なデータを区別（ここでは，中心に集まっているものが正常としている）

訓練データ



ここまでのまとめ

- **クラスタリングの基本**
  - 教師なし学習の一種である.
  - データポイントを似たグループ（クラスタ）に分類する手法.
  - クラスタ内のデータは互いに近く，異なるクラスタのデータは離れている.
- **クラスタリングの応用**

顧客データ分析，マーケティング戦略立案，Eコマースサイトでの商品カテゴリー自動分類

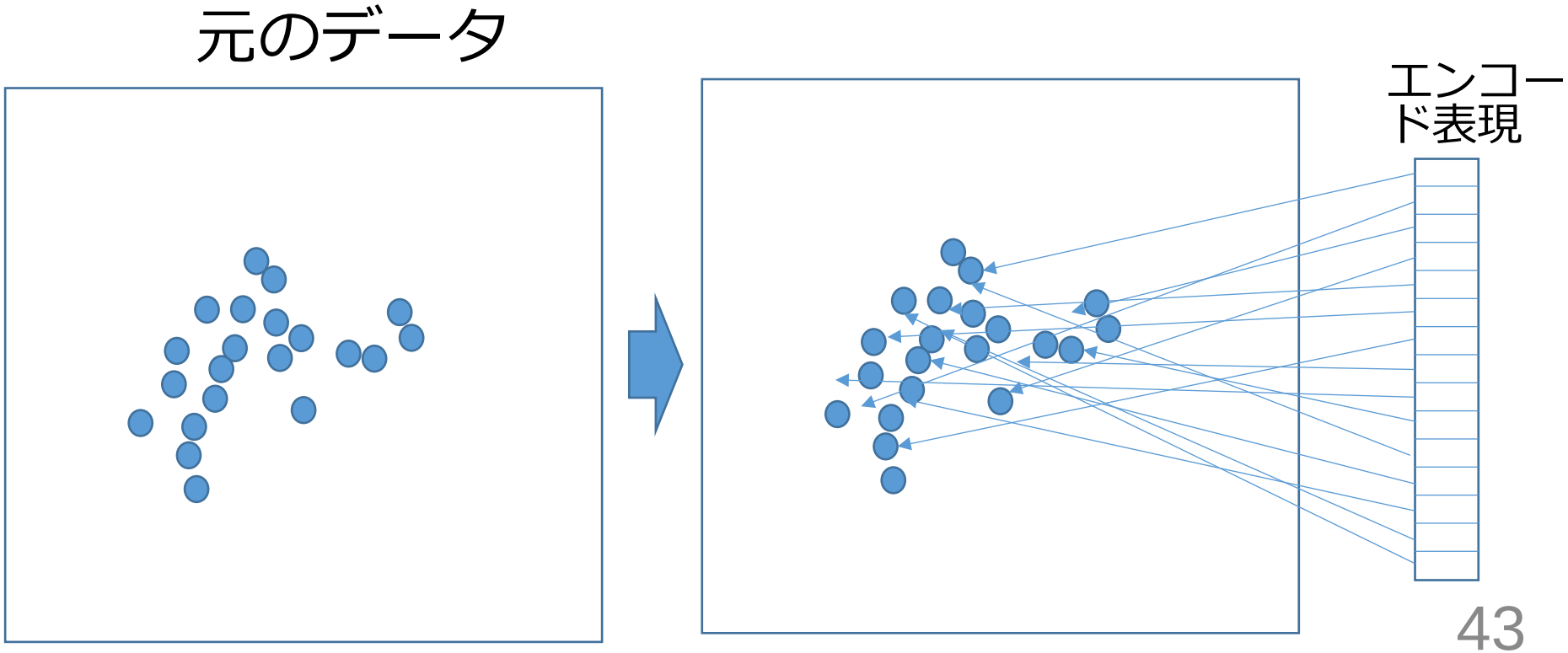
- K-means法の基本:
  - クラスタ数を指定して実行する
  - ランダムにクラスタ中心点を決め，データポイントを分類していく

4-4. オートエンコーダ

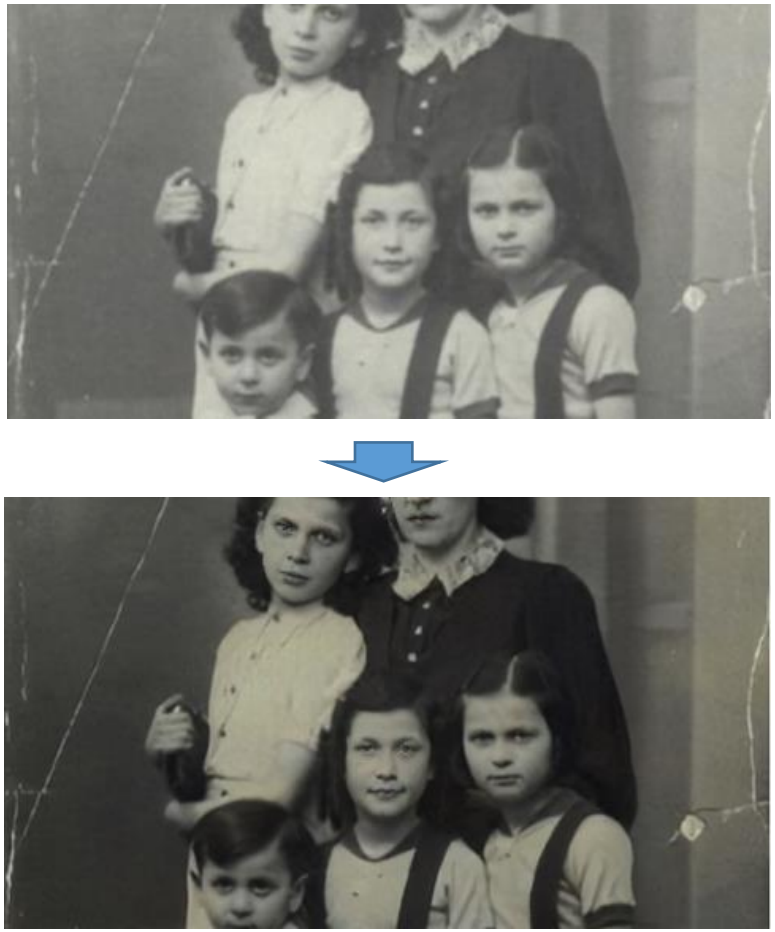
オートエンコーダ

- 元のデータを圧縮し，再構成する．
- そのとき，**元のデータの最も重要な特徴を自動的に学習し，元のデータの「エンコード表現」**を得る．

オートエンコーダは教師なし学習の一種



オートエンコーダの応用例



画像復元

写真からの顔の３次元化

オートエンコーダの主な応用分野

- **ノイズ除去**
  （例：古い写真や音声録音のクリーニング）
- **異常検知**
  （例：製造ラインでの不良品検出、正常データとの再構成誤差を利用）
- **情報検索**（例：圧縮された特徴を使って類似性を計算）
- **画像生成**（例：顔画像生成）

オートエンコーダの圧縮と再構成

元のデータ

８個の数値

１０１０１０１０

エンコード表現

元のデータを**圧縮**

0.993 0.838 0.995

再構成データ

８個の数値を再構成可能（完ぺきではない）

0.104 0.997 0.999,
0.968 0.997 0.945,
0.975 0.000

オートエンコーダの圧縮と再構成

１．データの圧縮と再構成

- 入力: [1, 0, 1, 0, 1, 0, 1, 0]
- 再構成: [0.998, 0.028, 0.979, 0.016, 0.985, 0.017, 0.975, 0.004]
- 高精度な再現: 1 → 0.9以上, 0 → 0.03未満

２．エンコード表現によるサイズ削減

- 元データは８つの数値 → エンコード表現は３つの数値
- エンコード例: [1, 0, 1, 0, 1, 0, 1, 0] → [0.993, 0.838, 0.995]
- 本質的特徴を保持しつつデータ圧縮

オートエンコーダによるノイズ除去

ノイズ付きのデータ

８個の数値

1 1 0 0 1 0 0 0

いくつかの数値が変化（ノイズ）

エンコード表現

ノイズ付きのデータを
圧縮し，再構成する

再構成データ

８個の数値

0.104 0.997 0.999
0.968 0.997 0.945
0.975 0.000

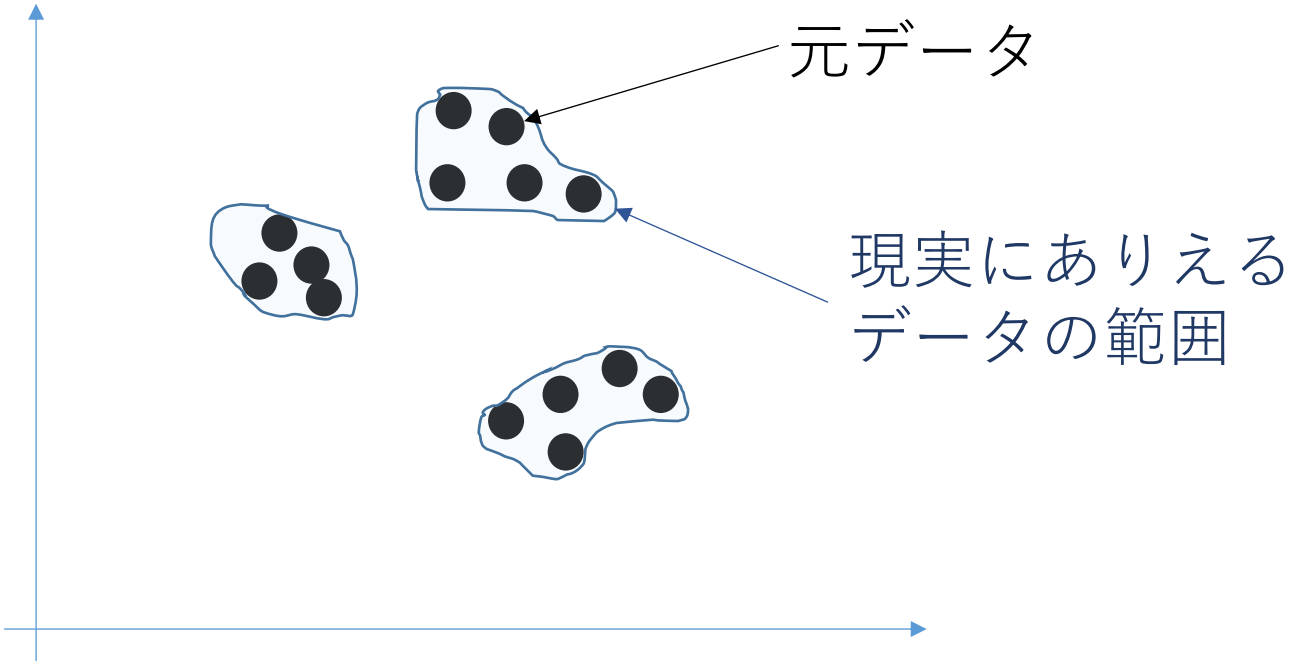
<u>再構成の結果</u>

・ノイズが除去され元データに近づくことが期待できる.

・学習に使用するデータ量を増やすことで，精度向上が期待できる

オートエンコーダによるデータ生成

オートエンコーダの結果として，**再構成（エンコード表現からの元データの再構成）ができる**ことから，「**現実にあり得るデータを生成できる能力を獲得**」と考えることもできる



オートエンコーダによるデータ生成

ランダムな数を 3つ入力

再構成データとして，

８個の数値

0.964 0.983 0.902 0.914 0.154
0.137 0.040 0.016

演習３．オートエンコーダ

ページ５１～５３

演習の目的

次の主要概念を理解

- 元のデータ
- 再構成データ
- エンコード表現
- ノイズ付きデータ
- データ生成（ランダムエンコーディングから生成された新しいパターン）

オートエンコーダ

① trinket の次のページを開く

https://trinket.io/python/23a86bc86ce8

② 実行結果が，次のように表示されることを確認

- 実行が開始しないときは、「**実行ボタン**」で**実行**
- ソースコードを**書き替えて再度実行**することも可能

オートエンコーダの種類

- **通常のオートエンコーダー：**基本的な構造、入力の再構成に焦点
- 変分オートエンコーダー (VAE)“潜在変数に確率分布を導入し、より柔軟なモデルを実現
- 積層オートエンコーダー：複数の層を重ねて、より複雑な特徴を学習
- 畳み込みオートエンコーダー：画像データに適した構造

オートエンコーダのまとめ

- **オートエンコーダの基本**
  - データ圧縮と再構成: **データの重要な特徴を保持し，効率的に再構成.**
  - **教師なし学習の一種: ラベルなし**で**特徴抽出**を行う.
- **オートエンコーダの機能**
  - **ノイズ除去**: データからノイズを削減し，元のパターンに近づける.
  - **データ生成**: ランダムな潜在表現から新しいデータを作成可能.
- **応用例**: ノイズ除去，異常検知，情報検索，画像生成。
- **オートエンコーダの種類**: 通常型，変分(VAE)，積層型，畳み込み型.

4-5. 画像生成AIとの関連

画像生成AIの基礎

教師なし学習の発展形であり，以下の特徴を持つ

- **複雑なパターンの学習と生成**
- **変分オートエンコーダー** (VAE)，**GANs**（敵対的生成ネットワーク：生成器と識別器が競争しながら学習）を使用
- 創造的なAI応用の基盤技術
- 応用例：

アート：スタイル変換、新しい芸術作品の創造

ファッション：新しいデザインの生成

製薬：新薬候補分子の設計

教師なし学習のまとめ

教師なし学習の基本

- 学習に**ラベルのないデータ**を使用

主なタスク

- **クラスタリング：似た特徴を持つデータをグループ化**
- **オートエンコード（自己符号化）**

- 学生生成AIに発展

4-5. 学習のバリエーション

自己学習のニュース

- **自己学習**は，**AI**が**自己対戦**や**他のAIとの対戦**を繰り返すことで，**その結果から学習を行う**機械学習の手法

AlphaZero のニュース

人工知能は，４時間の学習により，それまでの最強のチェスのコンピュータプログラムに勝利 (2017年)

https://www.theguardian.com/technology/2017/dec/07/alphazero-google-deepmind-ai-beats-champion-program-teaching-itself-to-play-four-hours より

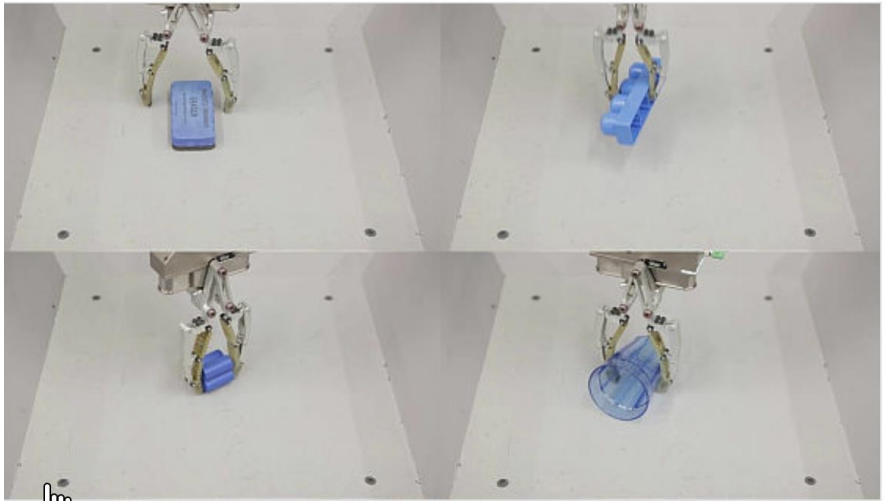
自己学習のニュース

OpenAIのニュース：
人工知能は、数ヶ月の学習により、オンラインゲーム Dota 2 で、人間のゲームチャンピョンに勝利。人工知能のプレイスタイルは、今後のゲームプレイの参考になるという見解もある（2018年）。

https://openai.com/blog/dota-2/ より

自己学習のニュース

人工知能がロボットを動かすことを多数繰り返し，学習を行う．（自力で上達する）



Google による，ロボットを用いた自己学習のニュース

　カメラ等を活用．ロボットハンドを動かして，物体を把持．ロボットハンドの動かし方，つかむ強さを自己学習し，さまざまな種類・大きさ・形状・素材に応じた把持を学習．（2016年）

https://www.youtube.com/watch?v=Q9tDHuidzak より

強化学習

複数の行動から
１つを選択

成功

失敗

失敗

行動は複数

成功は報酬大、
失敗は報酬小として
学習を行う

強化学習

強化学習は、**フィードバック（報酬やペナルティ）に基づく**学習.

使用するデータ

環境との相互作用を通じてデータを収集.

強化学習の主なメリット

- 長期的最適化： 即時の報酬だけでなく，長期的なフィードバックをもとに学習
- 動的環境対応： 環境の変化に適応する

強化学習の用途

ゲーム、ロボット、対話AIなどさまざま

学習のバリエーションのまとめ

- **自己学習**：AIが対戦を通じて追加データなしで学習（例：AlphaZero、OpenAI）
- **ロボティクスでの自己学習**：物体把持など複雑なタスクを学習
- **強化学習**：フィードバックに基づく学習，長期的最適化，動的環境対応
- **AI技術の急速な進歩**：人間の能力を超える事例が増加