# データサイエンス入門B

## ガイダンスと導入

児玉靖司

# もくじ

# データサイエンスの活用

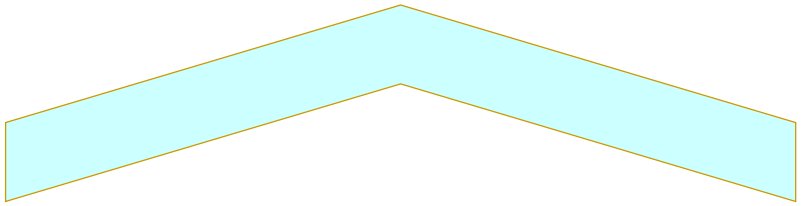
- データサイエンスとは=> データサイエンスA

  データを収集し、情報に変換し分析して、問題を解決する。
  「サイエンス（科学）」であるため、再現性も重要である。

- どんな問題を解決するか。
  - 顔認証、画像診断（医療）、株価予測、自動運転
  - 会計・金融
  - ファイナンス
  - マーケティング
  - 植物・医療
  - 音声処理

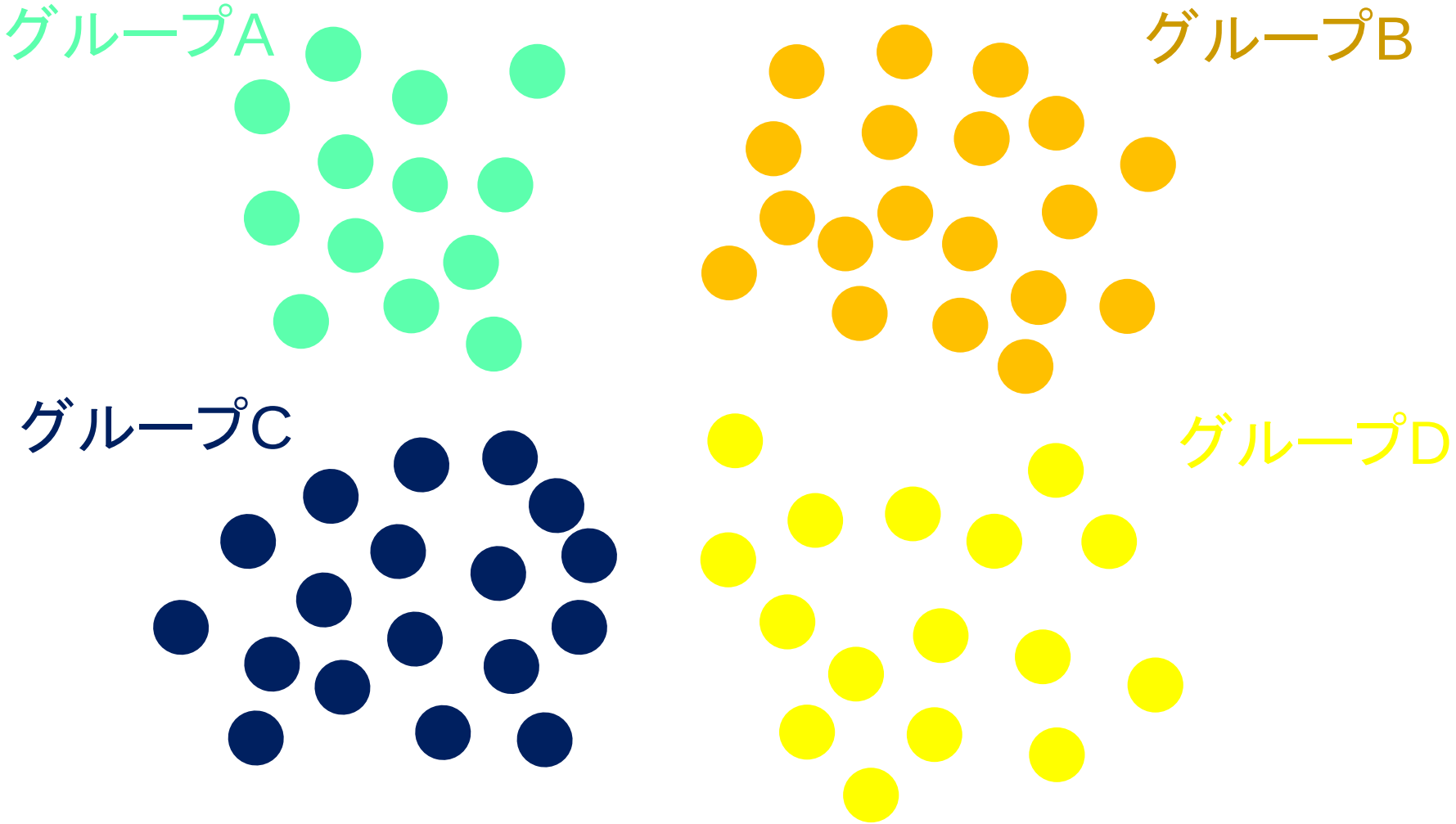
# データサイエンスとは

- どんな種類があるか。
  - 分類（クラスタリング） (clustering)
  - マッチング (matching)
  - 時系列解析 (time series analysis)

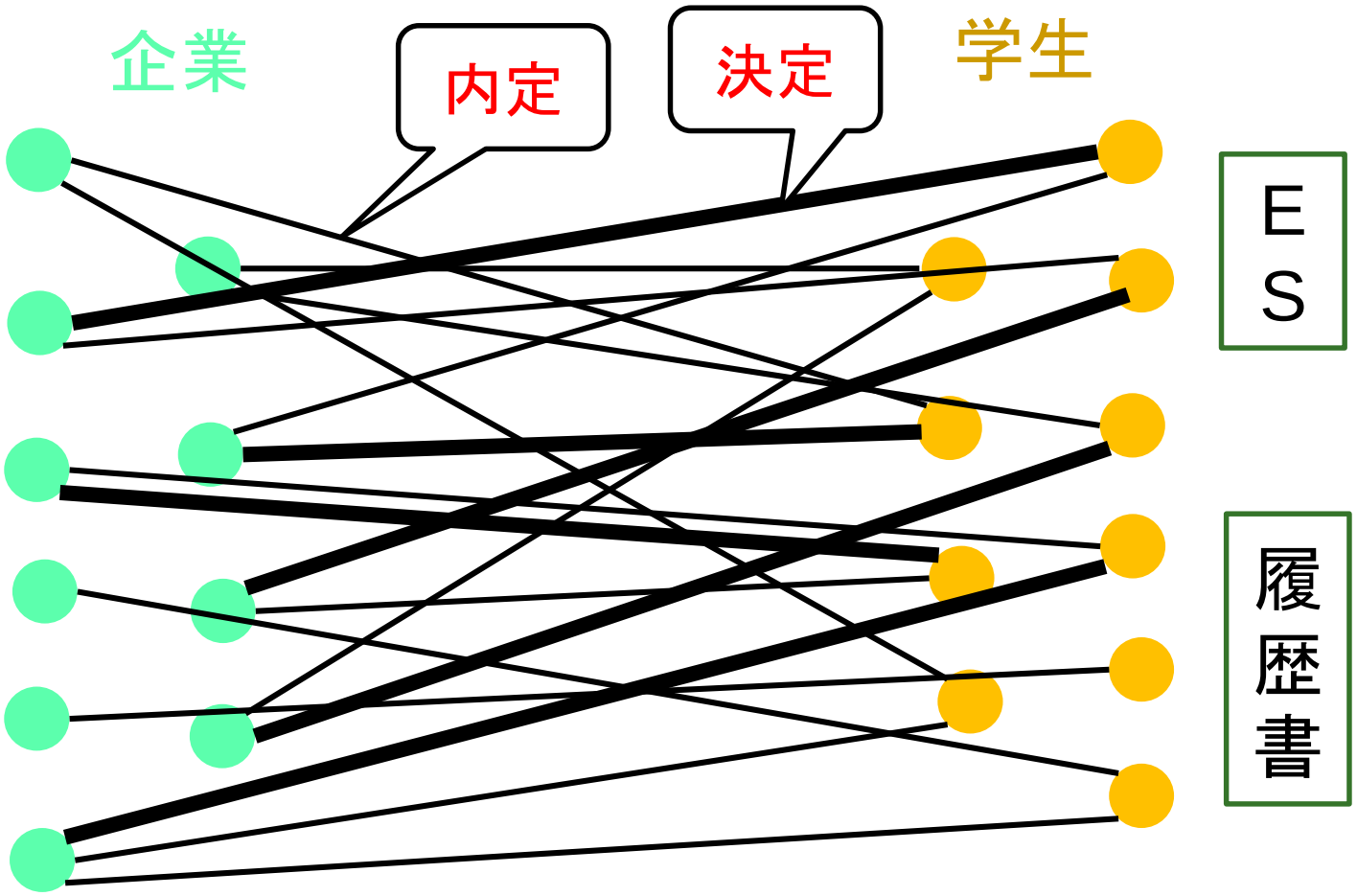


統計学、AI（人工知能）

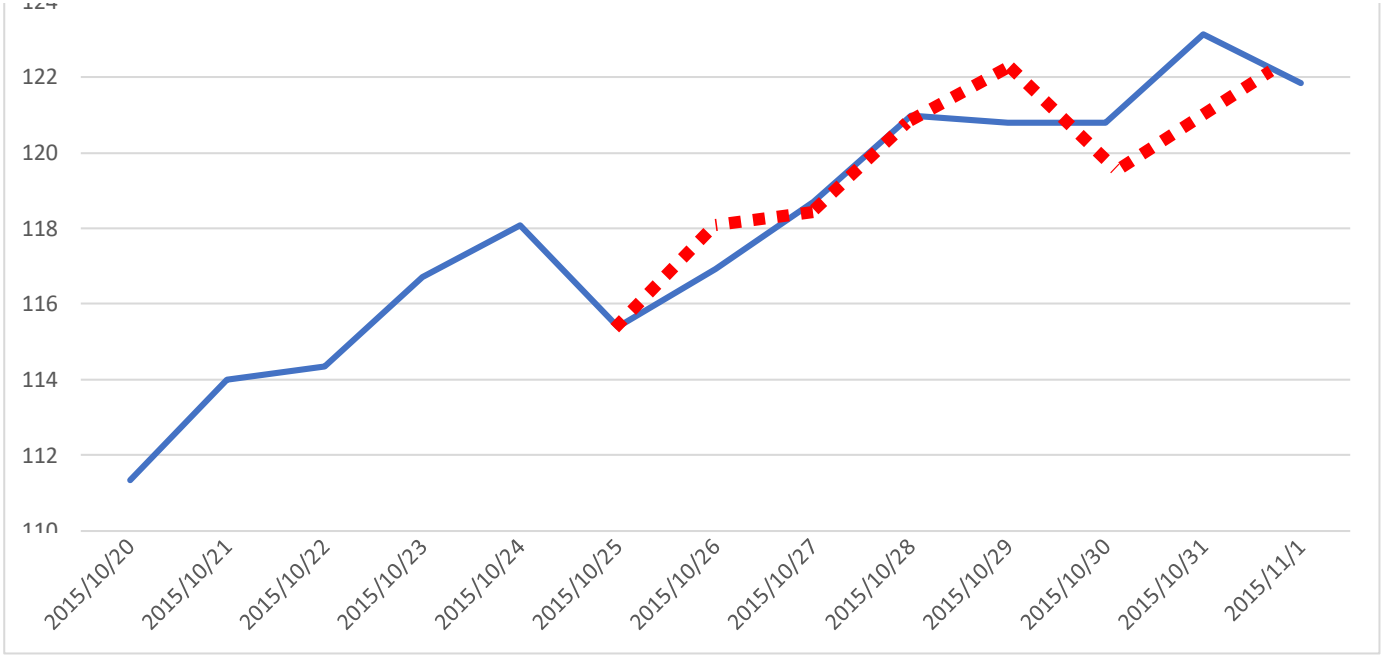
# 分類（クラスタリング、clustering）



# マッチング（matching）



# 時系列解析（time series analysis）




| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 111.34 | 114 | 114.33 | 116.7 | 118.08 | 115.4 | | | | | | |
| | 114 | 114.33 | 116.7 | 118.08 | 115.4 | 118.1 | | | | | |
| | | 114.33 | 116.7 | 118.08 | 115.4 | 118.1 | 118.4 | | | | |
| | | | 116.7 | 118.08 | 115.4 | 118.1 | 118.4 | 120.8 | | | |
| | | | | 118.08 | 115.4 | 118.1 | 118.4 | 120.8 | 122.2 | | |
| | | | | | 115.4 | 118.1 | 118.4 | 120.8 | 122.2 | 123.1 | |
| | | | | | | 118.1 | 118.4 | 120.8 | 122.2 | 123.1 | 191.1 |

# ビックデータと少ないデータ

1. **統計学**

   できるだけ少ない数のデータから全体のモデルを予測する。

2. **AI（機械学習、深層学習）**

   できるだけ多くの数（ビックデータ）のデータを使って分析をする。

これまでの**統計学**に加えて、2010年代からのAIの急速な発展により新しい分析法が生まれた。

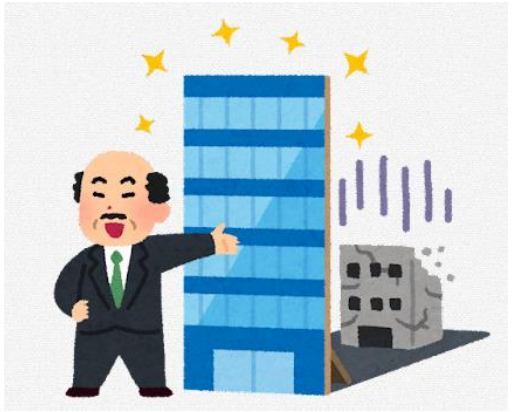
# 本講義での授業

2. 事例紹介１：会計・金融におけるデータサイエンス・AIの活用
3. 事例紹介２：ファイナンスにおけるデータサイエンス・AIの活用
4. 事例紹介３：マーケティングにおけるデータサイエンス・AIの活用
5. 事例紹介４：植物・医療におけるデータサイエンス・AIの活用
6. 事例紹介５：音声処理におけるデータサイエンス・AIの活用
7. e-Stat の使い方
8. データの分析１：ヒストグラム・箱ひげ図
9. データの分析２：平均・分散・標準偏差
10. データの分析３：散布図と相関係数
11. データの分析４：回帰分析
12. データ表現（可視化）
13. 質的なデータの解析
14. まとめ

# 第2回 会計・金融におけるデータサイエンス・AIの活用

経営学部：坂上学

- 金融詐欺(Financial Fraud)の検出
- 粉飾決算(Accounting Fraud)の検出
- 倒産予知(Bankruptcy Prediction)



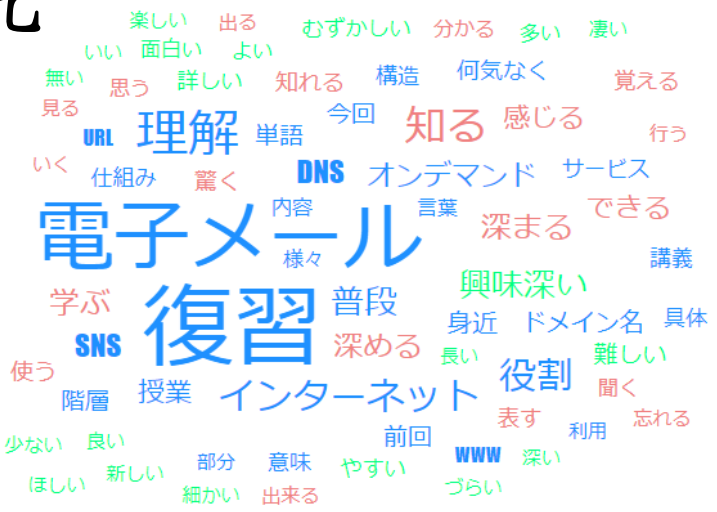
機械学習

# 第2回 会計・金融におけるデータサイエンス・AIの活用

経営学部：坂上学

- 財務数値だけでなく、有価証券報告書の文書にある記述についてテキストマイニングの手法を組み合わせて、倒産予知を行った研究
- テキストマイニング

  の基本的な手法




ワードクラウド

# 第3回 ファイナンスにおけるデータサイエンス・AIの活用

## 機械はファイナンスを学習できるか?

経営学部：高橋慎

機械は何でもできるように見えるが…

- 音声認識、翻訳、戦略ゲーム、運転、物体操作


ファイナンスは学習できるのか？

- ファイナンスは**別物**（変化する市場、競争、認知バイアス…）

それでも…機械学習への期待は大きい

- The Economist「機械学習が金融の大部分を変えると期待」

  https://www.economist.com/finance-and-economics/2017/05/25/machine-learning-promises-to-shake-up-large-swathes-of-finance

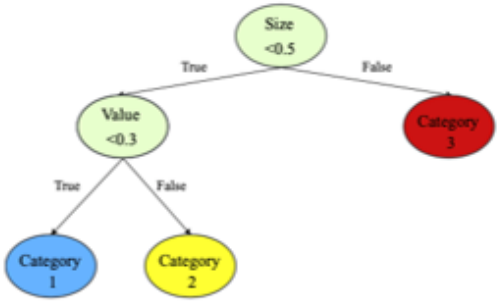
# 第3回 ファイナンスにおけるデータサイエンス・AIの活用

## 機械はファイナンスを学習できるか?

ファイナンス分野での応用研究例

経営学部：高橋慎

1. リターン（株価の変化率）予測
   - 株式銘柄選択における機械学習モデルの比較分析
     https://doi.org/10.1093/rfs/hhaa009




2. 自然言語処理の活用
   - ビジネスニュースからの情報抽出と株式銘柄選択への活用
     https://ssrn.com/abstract=3389884


# 第4回 マーケティングにおけるデータサイエンス・AIの活用

経営学部：長谷川翔平

- マーケティングとは？
  - STPマーケティング
  - マーケティング・ミックス
- マーケティング・データ
  - 伝統的なデータ：小売店での購買履歴
  - 新しいデータ：ECサイトでの購買履歴，レビュー，SNS，etc.
- データの活用
  - 売上予測
  - マーケティングの最適化・効果測定

# データサイエンス入門B　第5回

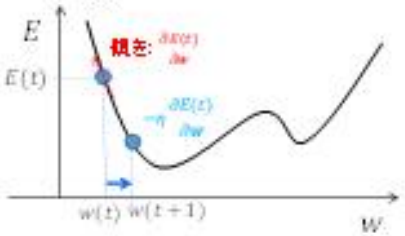
理工学部　応用情報工学科　彌冨　仁
iyatomi@hosei.ac.jp

## 実例紹介4　植物・医療におけるデータサイエンス・AIの活用

導入：　人工知能(Artificial Intelligence)と
深層学習(Deep Learning)


機械の"学習"とは？　なぜComputerは賢くなれるのか？

重要なDeep Learning model – CNNとGAN


応用事例の紹介:

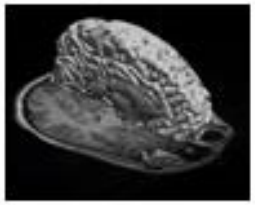
植物病害の自動診断技術と実用化の紹介

医療分野の応用事例の紹介　など

植物病害自動診断技術の開発
農水省受託プロジェクト

自習のためのおすすめ図書紹介:

皮膚がんの自動診断



脳MRIを対象とした
類似症例検索

胃がんの自動診断

# 第6回 音声処理におけるデータサイエンス・AIの活用

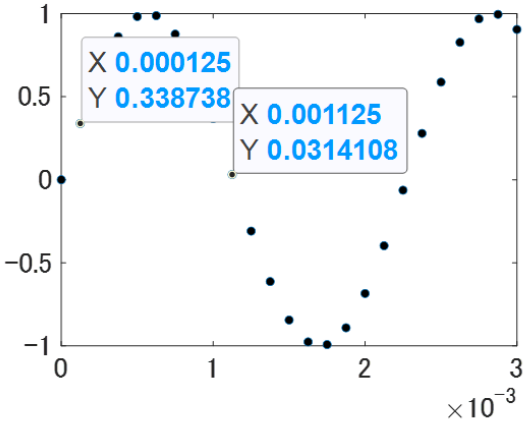
情報科学部：伊藤克亘

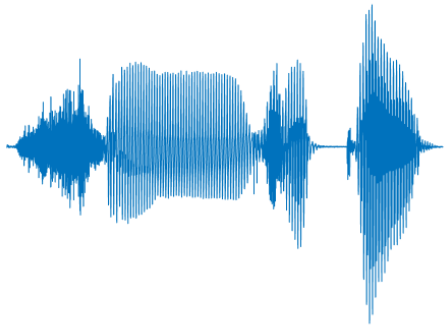
1）コンピュータによる音の処理

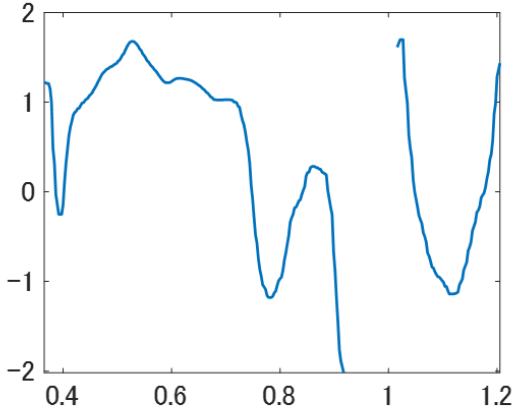
音声認識　音声合成　音声対話　音声強調

音楽制作　音声分析　…

2）データとしての音



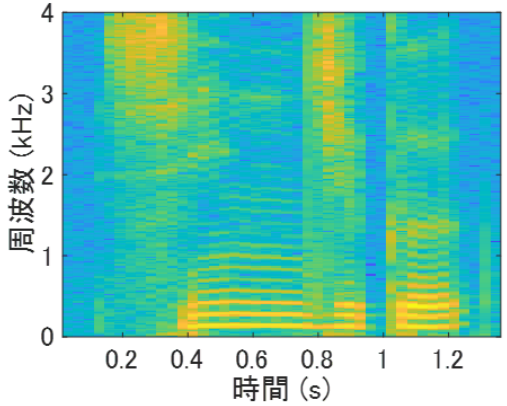



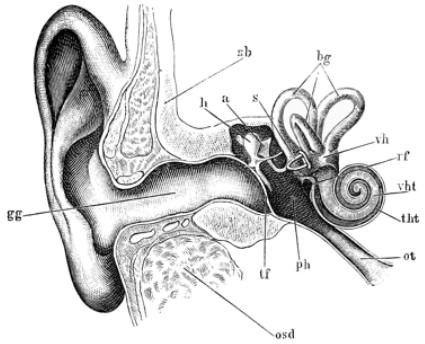


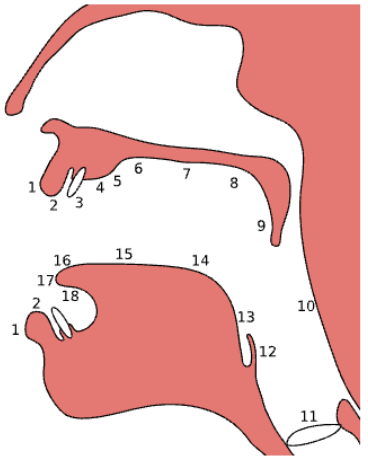
# 第6回 音声処理におけるデータサイエンス・AIの活用

情報科学部：伊藤克亘

3）音のデータサイエンス





4）ことばのデータサイエンス

| | 今日 | 明日 | 雨 | 晴れ | 寒い |
|---|---|---|---|---|---|
| 今日は雨です | 1 | 0 | 1 | 0 | 0 |
| 明日は晴れますか | 0 | 1 | 0 | 1 | 0 |
| 今日は晴れです | 1 | 0 | 0 | 1 | 0 |
| 今日は寒いです | 1 | 0 | 0 | 0 | 1 |

# 第7回　e-Statの使い方

演習があります
各自で **Excel** を利用できるPC環境を 用意すること
（**Windows, Mac**）

- Excelの準備

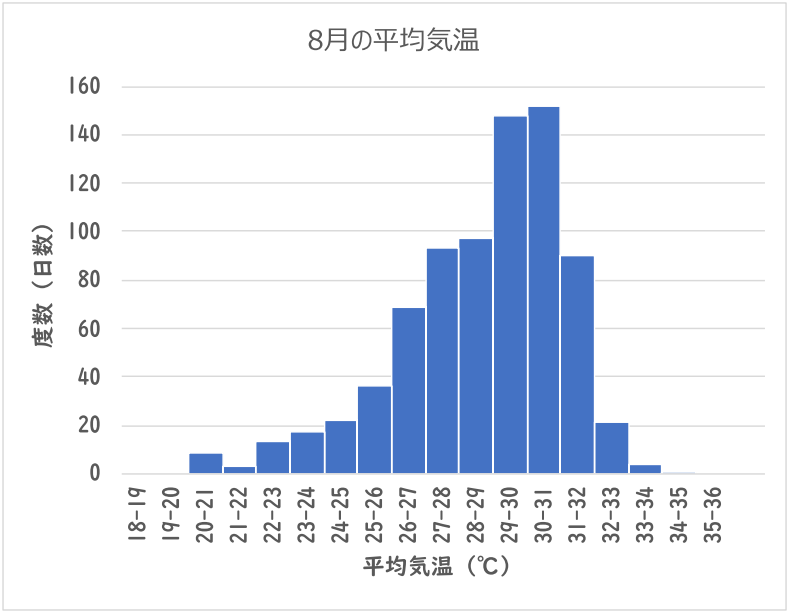
- Excelを利用した統計データ分析
- e-Statの使い方

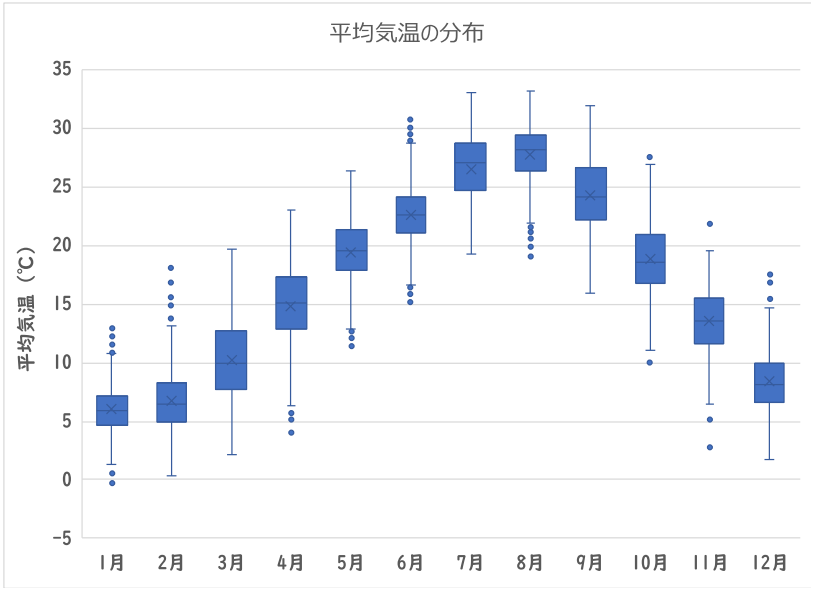

https://www.e-stat.go.jp/

# 第8回　ヒストグラム・箱ひげ図

- データの種類
- 度数分布の可視化

演習があります
各自で **Excel** を利用できるPC環境を 用意すること
（**Windows, Mac**）



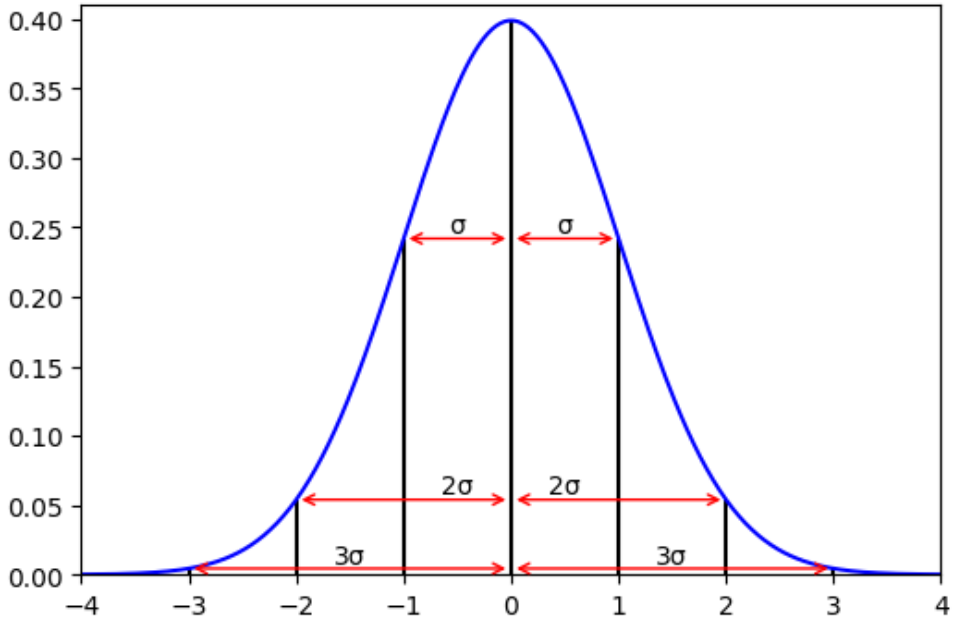




https://www.jma.go.jp/jma/

# 第9回　平均・分散・標準偏差

- 確率変数の期待値と分散
- 正規分布

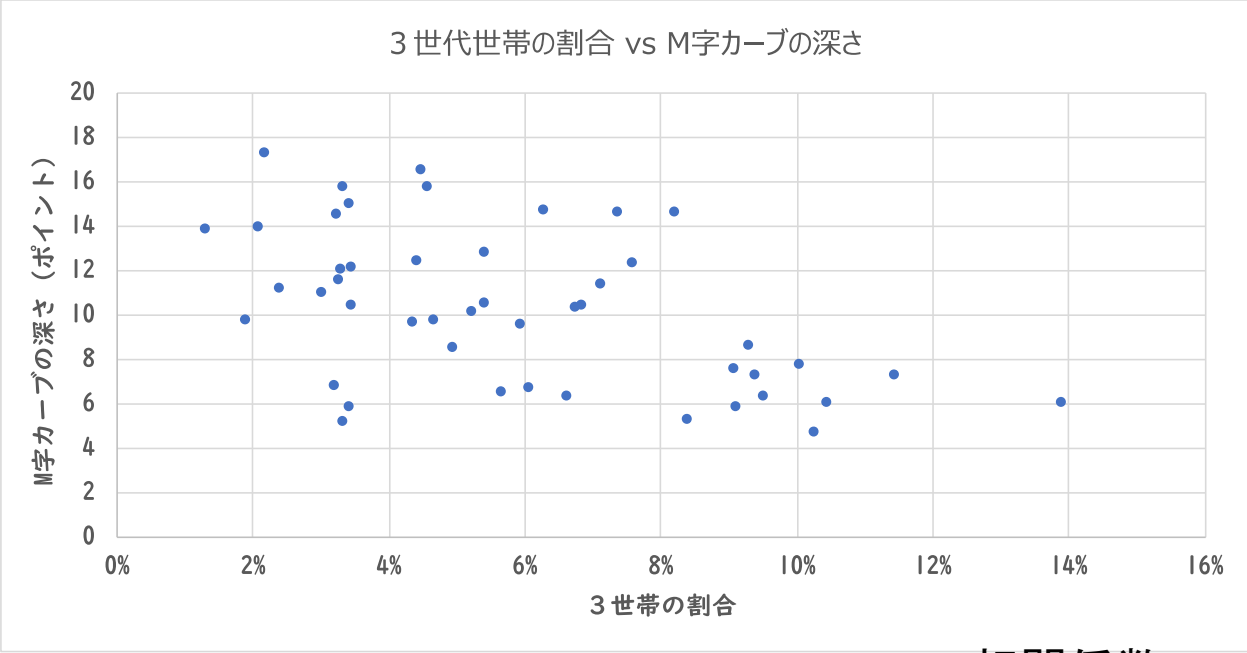
演習があります
各自で **Excel** を利用できるPC環境を 用意すること
（**Windows, Mac**）



# 第10回　散布図と相関係数

- ２変数の関係
- 因果関係

演習があります
各自で **Excel** を利用できるPC環境を 用意すること
（**Windows, Mac**）

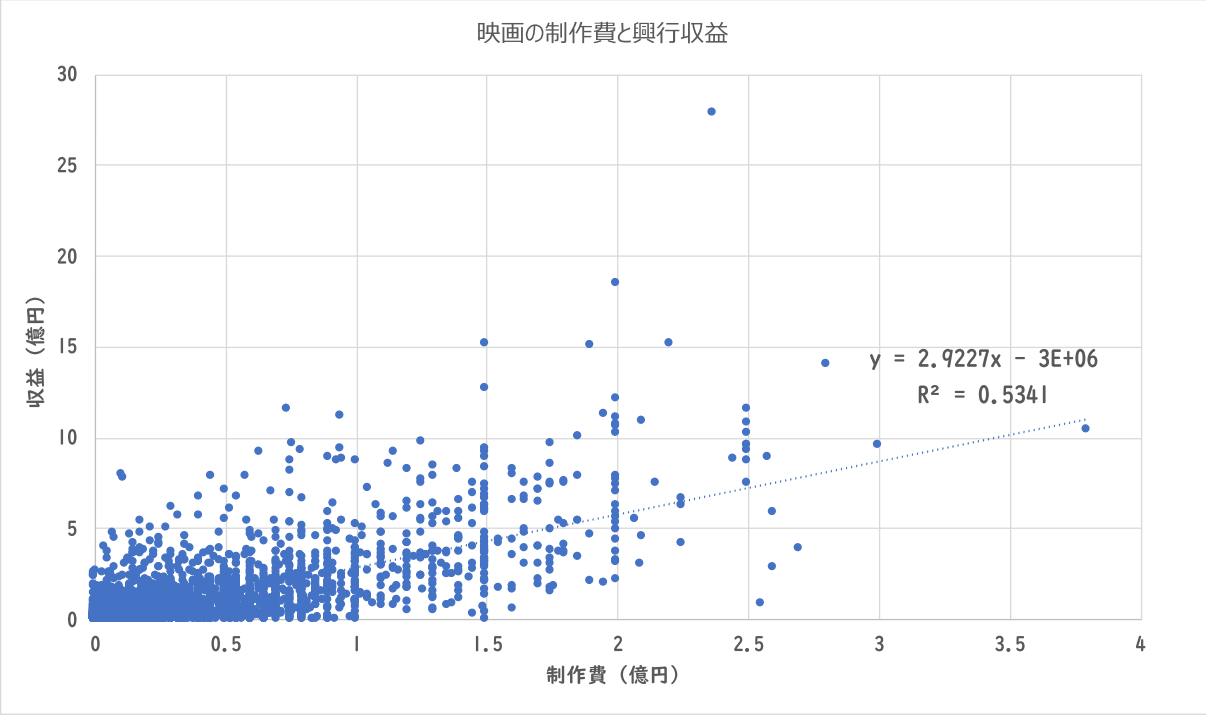



相関係数：-0.52

# 第11回　回帰分析

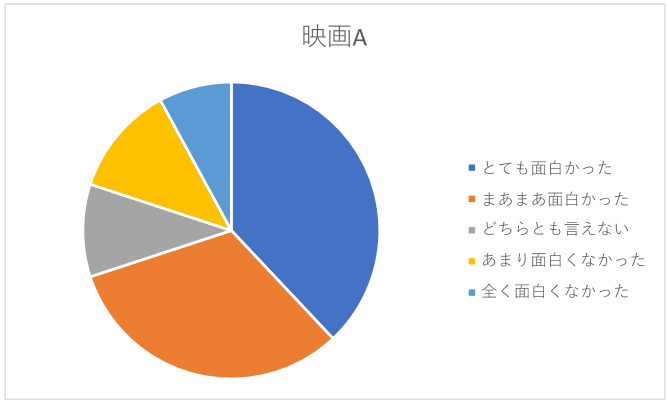
- 単回帰
- 重回帰
- 主成分分析

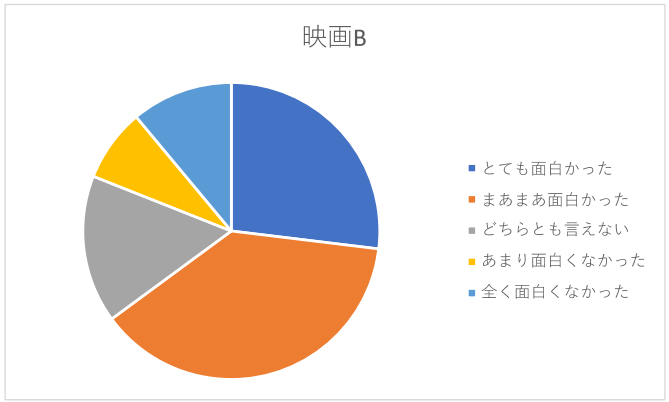
演習があります
各自で **Excel** を利用できるPC環境を 用意すること
（**Windows, Mac**）

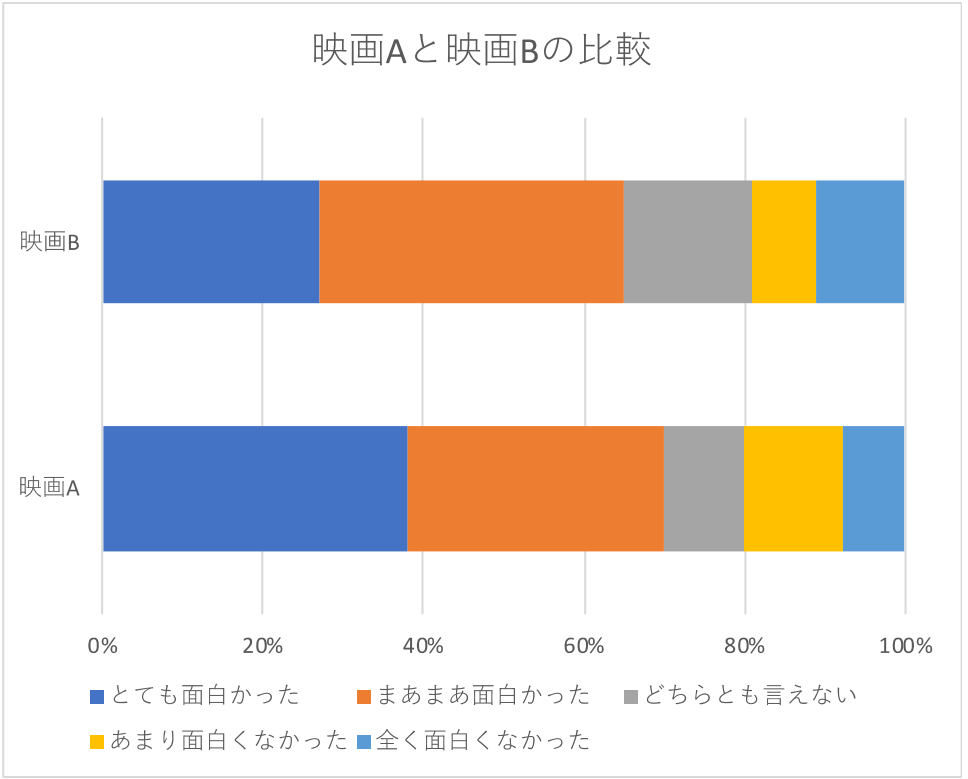


# 第12回　データ表現（可視化）

## 判定・意思決定のため、目的に合った可視化を

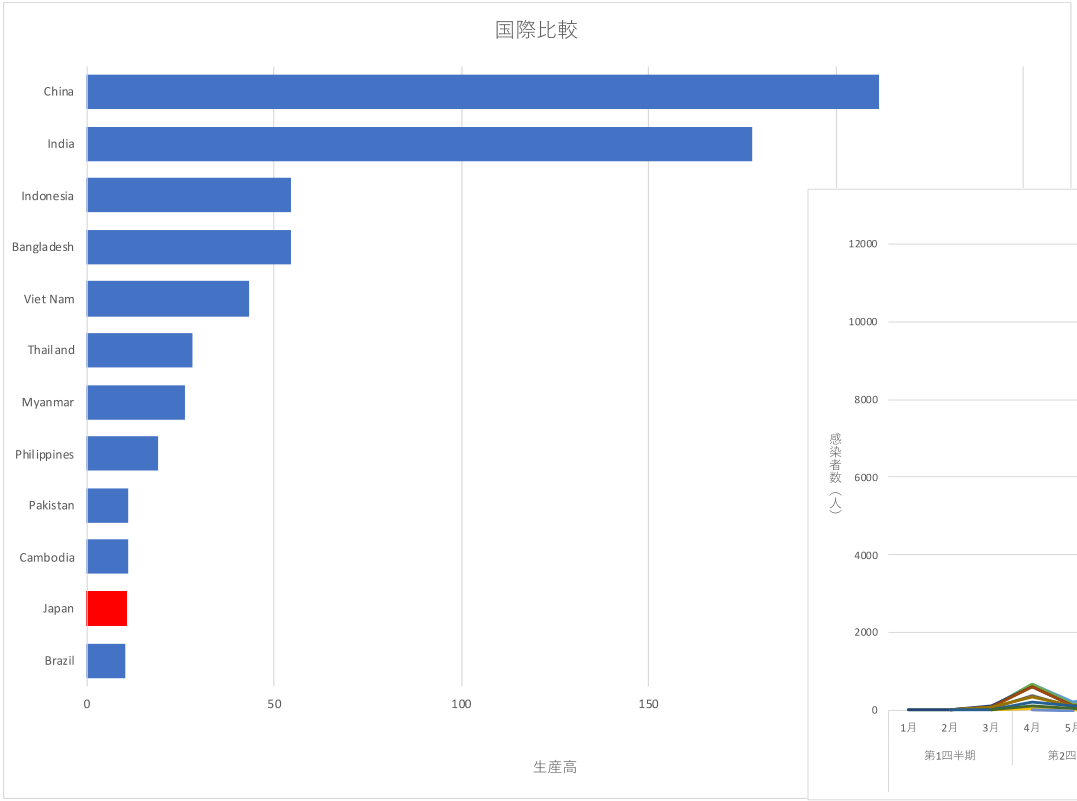


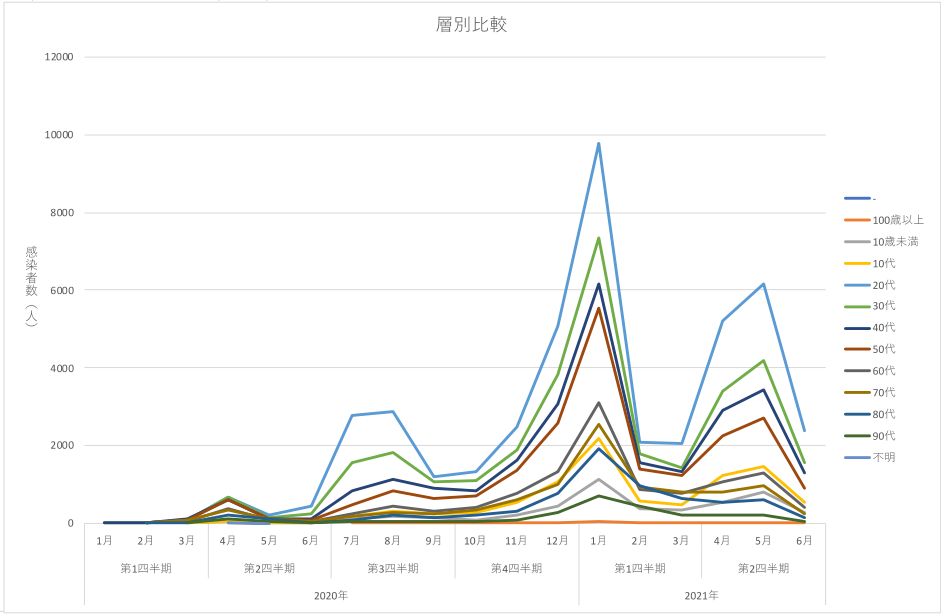



# 第12回　データ表現（可視化）

判定・意思決定のため、目的に合った可視化を

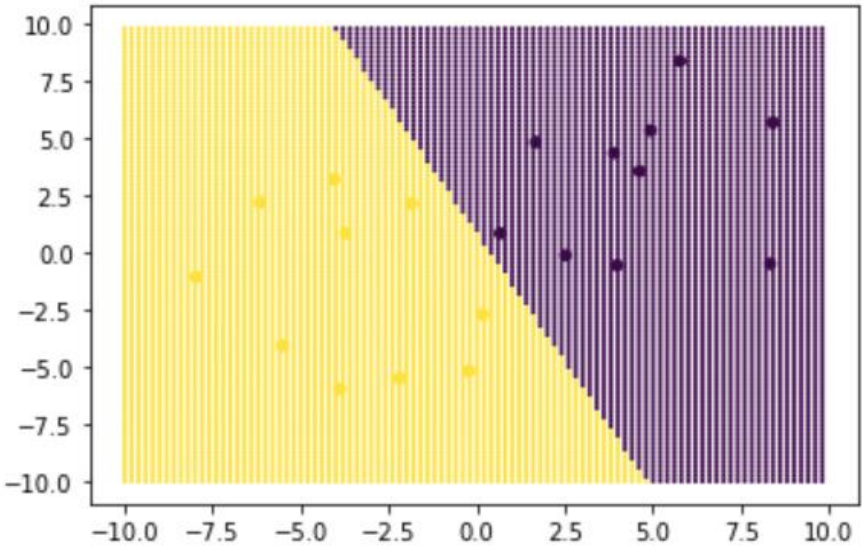


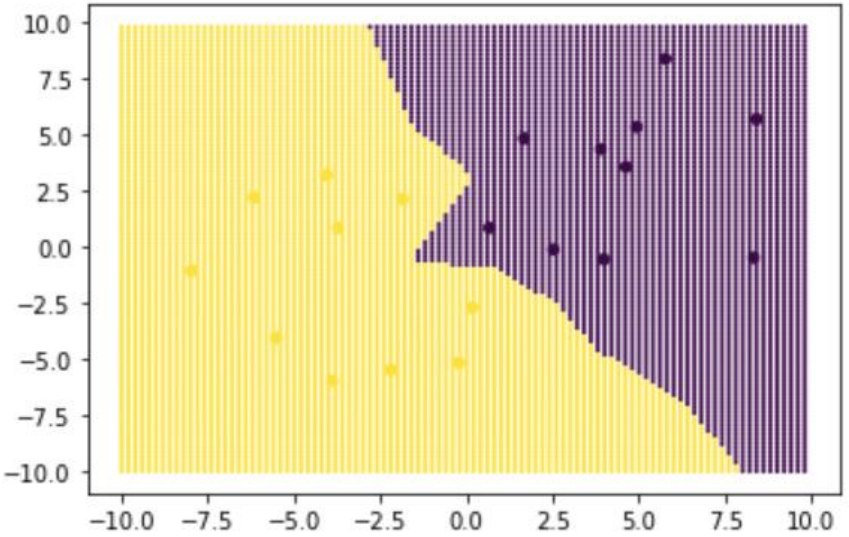



# 第12回　データ表現（可視化）

機械学習で分類

同じ学習データでも分類器の種類によって結果が異なる



# 第13回　質的なデータの解析

アンケート調査結果などの質的データ

選択肢　→　コード化

| No. | 映画 | アンケート |
|---|---|---|
| 1 | 映画A | 満足 |
| 2 | 映画A | 不満 |
| 3 | 映画B | 満足 |
| 4 | 映画A | 満足 |
| 5 | 映画B | 不満 |
| : | : | : |

| No. | 映画 | アンケート |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 0 | 0 |
| 3 | 1 | 1 |
| 4 | 0 | 1 |
| 5 | 1 | 0 |
| : | : | : |

# 第13回　質的なデータの解析

クロス集計･･･どちらが満足しているか

| | 満足 | 不満 | 合計 |
|---|---|---|---|
| 映画A | 291 | 125 | 416 |
| 映画B | 270 | 146 | 416 |
| 合計 | 561 | 271 | 832 |

行パーセント

| | 満足 | 不満 | 合計 |
|---|---|---|---|
| 映画A | 0.70 | 0.30 | 1.00 |
| 映画B | 0.65 | 0.35 | 1.00 |
| 合計 | 0.67 | 0.33 | 1.00 |

列パーセント

| | 満足 | 不満 | 合計 |
|---|---|---|---|
| 映画A | 0.52 | 0.46 | 0.50 |
| 映画B | 0.48 | 0.54 | 0.50 |
| 合計 | 1.00 | 1.00 | 1.00 |

# 第14回 まとめ

経営学部：児玉靖司

- データサイエンスの活用
- データサイエンスとは
- 事例紹介
- e-Stat の使い方、データの分析
- データ表現
- 質的なデータの解析