情報Ⅱ オンライン学習会

～統計的手法によるデータ分析～

早稲田大学　蓮池隆

本日の学習会では…

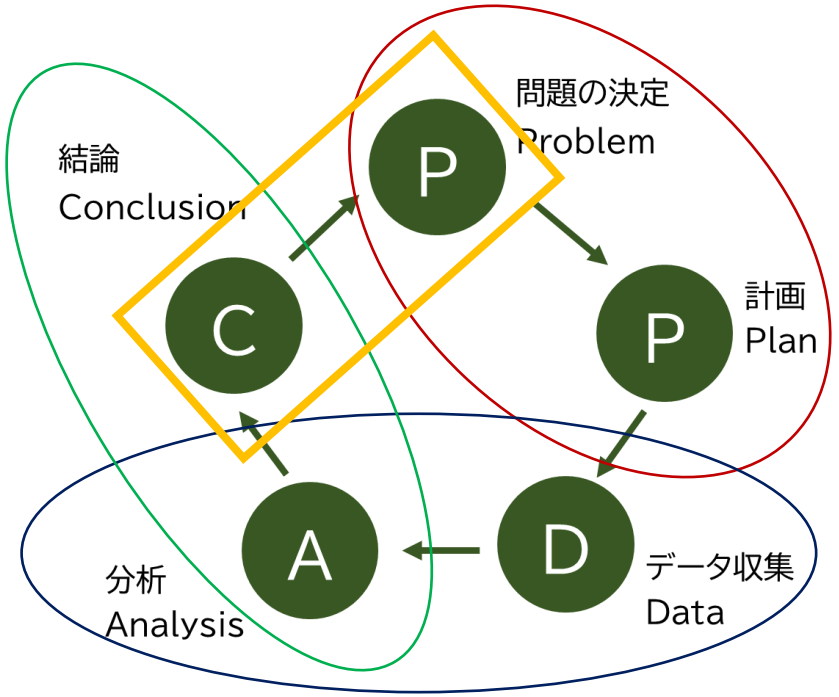
- 本日の学習会では，今後公開予定の情報IIに関する講義動画
  - 情報とデータサイエンス(1)：重回帰分析を用いた予測『睡眠時間を他の行動時間から予測しよう』
  - 情報とデータサイエンス(2)：主成分分析による次元縮約『データを圧縮して、関係を見よう！』
  - 情報とデータサイエンス(4)：クラスタリング『自分と近い性格の人は誰？』

  を基に，ポイントを解説していきます.

データサイエンスとは…

- **データサイエンス**とは
  - さまざまな課題の解決や展望を予測するため，
  - 膨大に蓄積されているデータの内容やその分布を調べ，
  - 特定の傾向や性質に基づいた解析により，適切な解決方法を提示・評価する

(日本大百科全書(ニッポニカ))



一度で課題をすべて
解決できるのは稀

結果から新たな課題
に取り組む

データサイエンスの定義は他にも…

- データを用いて新たな科学的および社会に有益な知見を引き出そうとするアプローチのことであり、その中でデータを扱う手法である情報科学、統計学、アルゴリズムなどを横断的に扱う (Wikipedia)

(教科書では)

- データを処理・分析し，データから価値を創造するための技術
- 社会で日々やり取りされるデータの種類や量が膨大となり(≒ビッグデータ)，このようなデータを利活用し，新たな価値を創造する，意味のある情報や規則性，関連性などを導き出す手法

- 他にも書籍やWebなどでたくさんの説明が書かれている

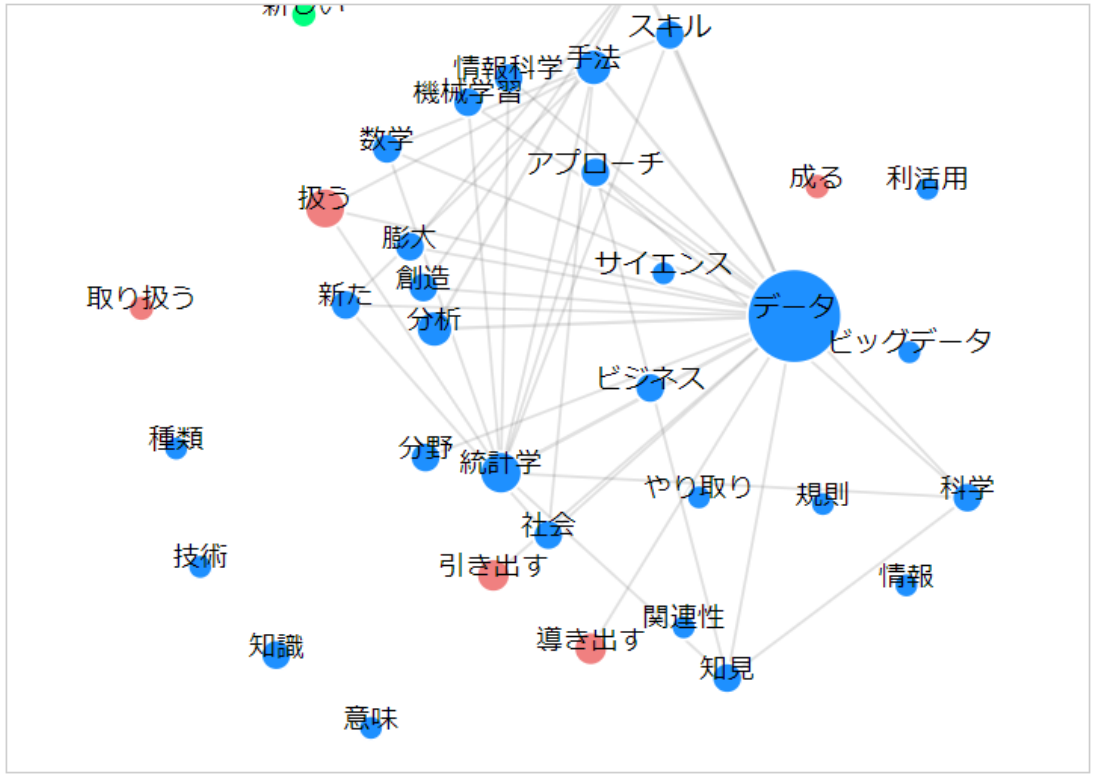
  → **キーとなる言葉は何でしょうか？**
  (あなたの主観ですよね？と言われないためには，データ分析が必要)

先ほどの文章を分析する

共起ネットワーク

共起回数をダウンロード(β)

文章中に出現する単語の出現パターンが似たものを線で結んだ図。出現数が多い語ほど大きく、また共起の程度が強いほど太い線で描画されます。



ワードクラウド

単語の出現頻度をダウンロード

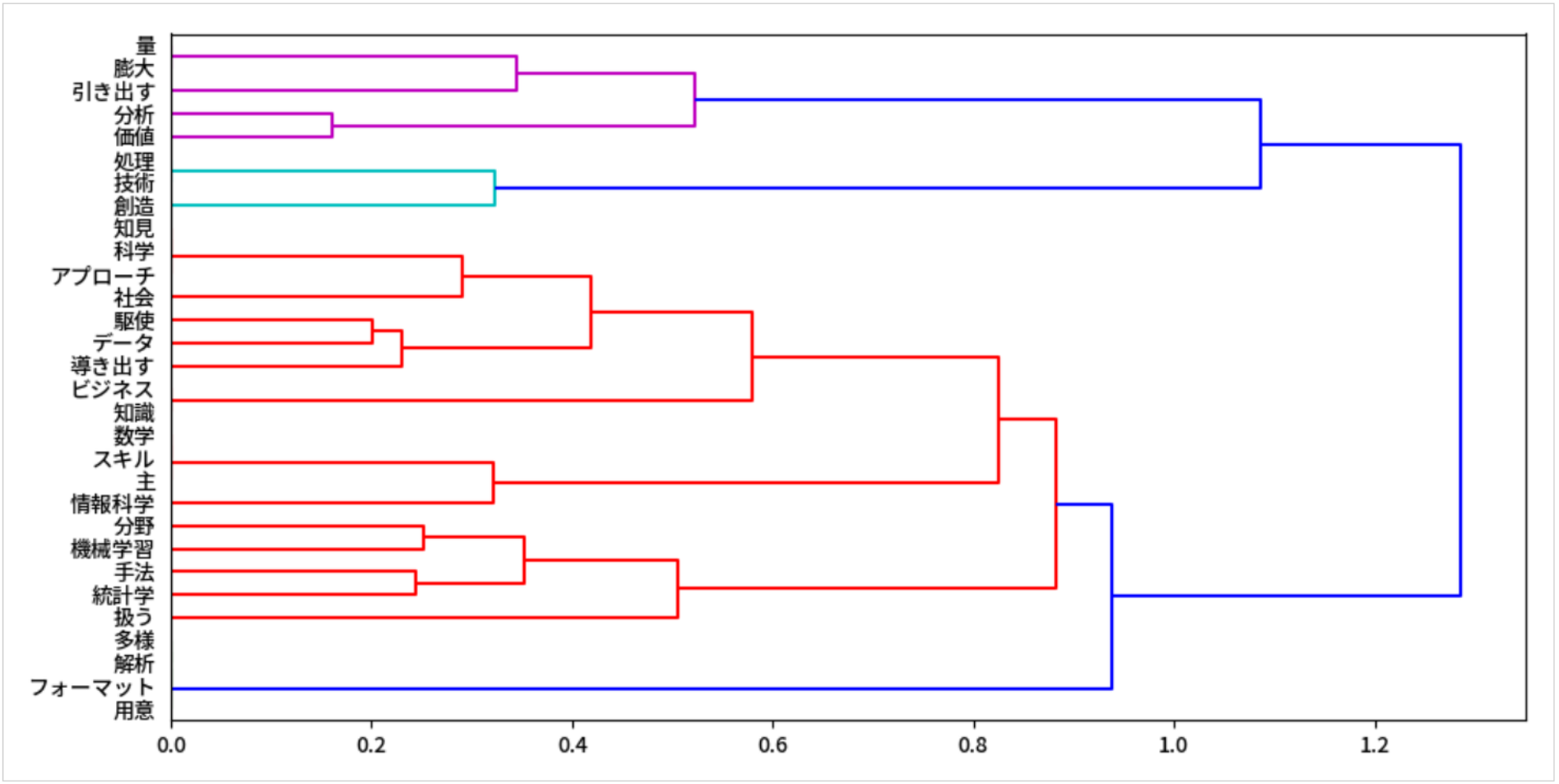
スコアが高い単語を複数選び出し、その値に応じた大きさで図示しています。色が品詞に対応しています。

(出典)　https://textmining1.userlocal.jp/home/result/129a35b46ac0638ece0c62f2c64bbc2a

先ほどの文章を分析する

階層的クラスタリング(β版)

文章中での出現傾向が似た単語をまとまりとしてとらえられるよう樹形図で表したものです。グループは色分けして表示しています。



この一連の流れは，まさしく「データサイエンス」という言葉をデータ分析している！

情報Iでのデータサイエンス

	「社会と情報」「情報の科学」	「情報Ⅰ」
統計	数学と連携して 平均値，中央値 などの基本的統計値を扱う	**分散，標準偏差，相関係数などの統計指標，散布図，仮説検定の考え方，交絡因子**なども扱う
分析	主にグラフ化などを行い，データの傾向をつかむ	**クロス集計，仮説検定，単回帰分析，これらを通じたデータの可視化，現象のモデル化と予測**
量的データ	主に表形式で整理された数値を中心に扱う	**量的データ**の記載あり。**表形式で整理されていないものも扱う**
質的データ	質的データの記載なし テキストマイニングの例あり	**質的データ**の記載あり テキストマイニングの例あり
扱うデータ	整理されたデータを扱う	実験値などの**整理されていないデータも扱い，外れ値，欠損値**などの処理も学ぶ
尺度	―	名義，順序，間隔，比例など**尺度水準の違い**を扱う
データベース	「情報の科学」のみで扱う	**情報を収集・蓄積・提供する方法として全員が学ぶ**

中学校数学科「Dデータの活用」，高校「数学Ⅰ」の(4)「データ分析」と連携
赤字＝数学科で学び情報科で活用　赤字＝情報科のみで活用

(出典)
https://www.sky-school-ict.net/shidoyoryo/210108/

(再掲)本日の学習会では…

- 本日の学習会では，今後公開予定の情報IIに関する講義動画
  - 情報とデータサイエンス(1)：重回帰分析を用いた予測『睡眠時間を他の行動時間から予測しよう』
  - 情報とデータサイエンス(2)：主成分分析による次元縮約『データを圧縮して、関係を見よう！』
  - 情報とデータサイエンス(4)：クラスタリング『自分と近い性格の人は誰？』

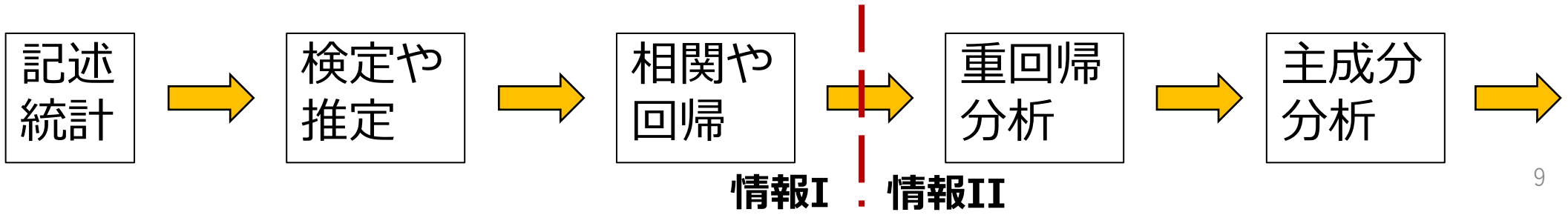
  を基に，ポイントを解説していきます.

- ポイント：「**統計的手法をデータ分析(何が重要？どれが効いてる？)に活かす**」
  - 予測したい・目的(結果)に対して何が効いてるか知りたい　→　**回帰分析**
  - 変数を減らして(統合して)うまく説明したい　→　**主成分分析**
  - データをうまく分類したい　→　**クラスタリング**
  - 仮説が正しいかチェックしたい　→　**検定(推定)**

情報Iから情報IIへ

- 情報Iで学習してきたこと(できること)
  - 収集データを観察する(欠損値，外れ値，ヒストグラムなどのグラフ表現，散布図など)
  - 基本統計量の計算(平均値，分散，相関係数)
  - 単回帰分析(説明変数が1つの線形回帰)
  - (一部，検定や推定)

- ここから先のデータ分析手法を学んでいくのが情報II
  - 回帰分析で説明変数を複数にしたい(重回帰分析)
  - 予測だけでなく分類や要約をしたい(主成分分析・クラスタリング)



分析するための言語・ソフトウェア

- PythonでもRでもOK！もちろんExcelでもできる！
- PythonやRを使えた方が，今後大学や企業などでデータ分析を行うときには便利(様々なライブラリが用意されている)

(例：Pythonで線形回帰を行うプログラム．詳細は講義動画参照)

```
1 from sklearn.linear_model import LinearRegression
2 model = LinearRegression()
3 model.fit(X, y)
4 print(model.intercept_, model.coef_)
```

(例：Rで主成分分析を行うプログラム．詳細は講義動画参照)

```
> result <- prcomp(syouhi,scale=T)
```

情報とデータサイエンス

重回帰分析を用いた予測

データに関しては総務省統計局「社会生活基本調査」から引用
https://www.stat.go.jp/data/shakai/2021/kekka.html

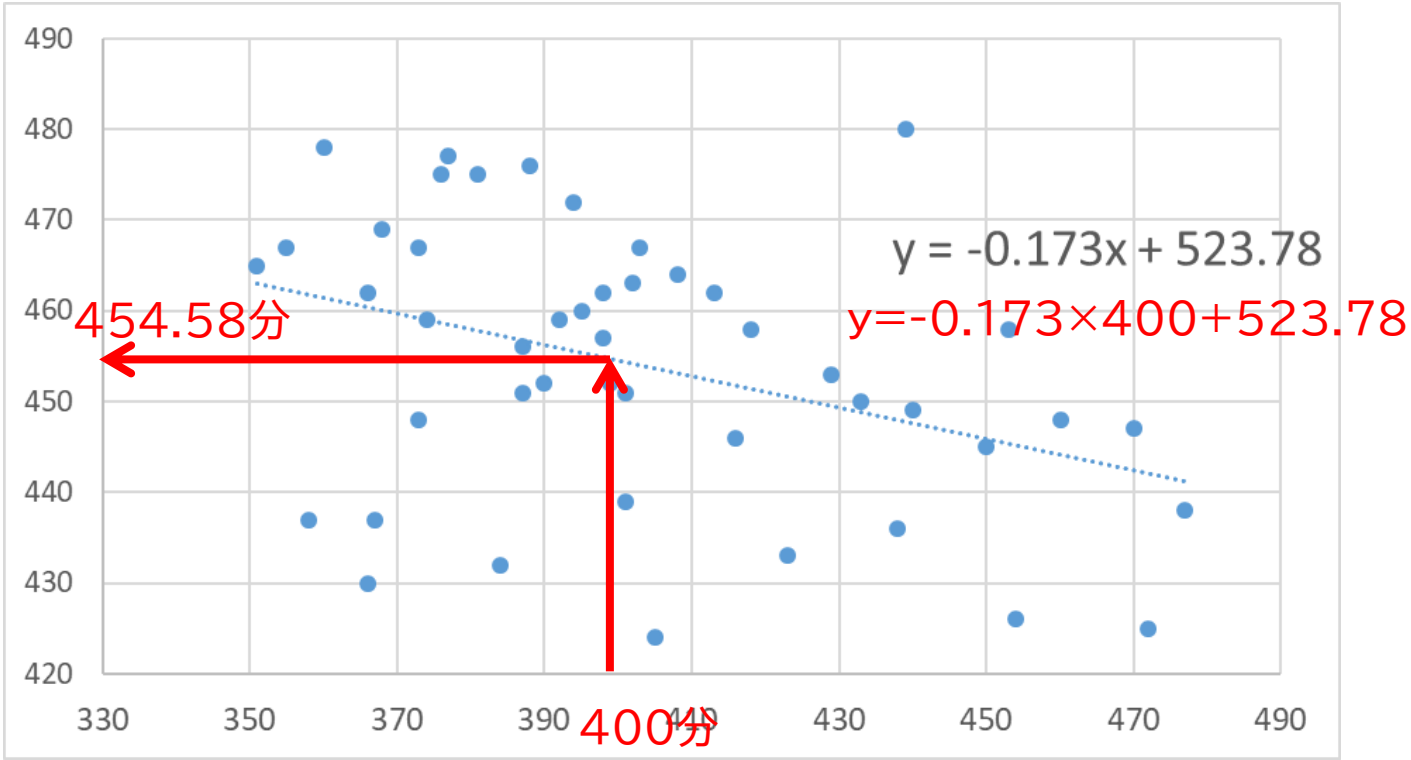
情報Ⅰで学んだこと

単回帰分析：睡眠時間（目的・結果）を学業時間から予測する

データ

	目的変数	説明変数
地域区分	睡眠	学業
01_北海道	467	355
02_青森県	469	368
03_岩手県	458	453
04_宮城県	448	373
05_秋田県	467	373
06_山形県	477	377
07_福島県	436	438
08_茨城県	456	387
09_栃木県	472	394

回帰直線 → 予測



学習時間以外のデータも使った方が・・・

睡眠時間を予測するために、
学業の時間だけでなく他の時間も使うことはできないか？

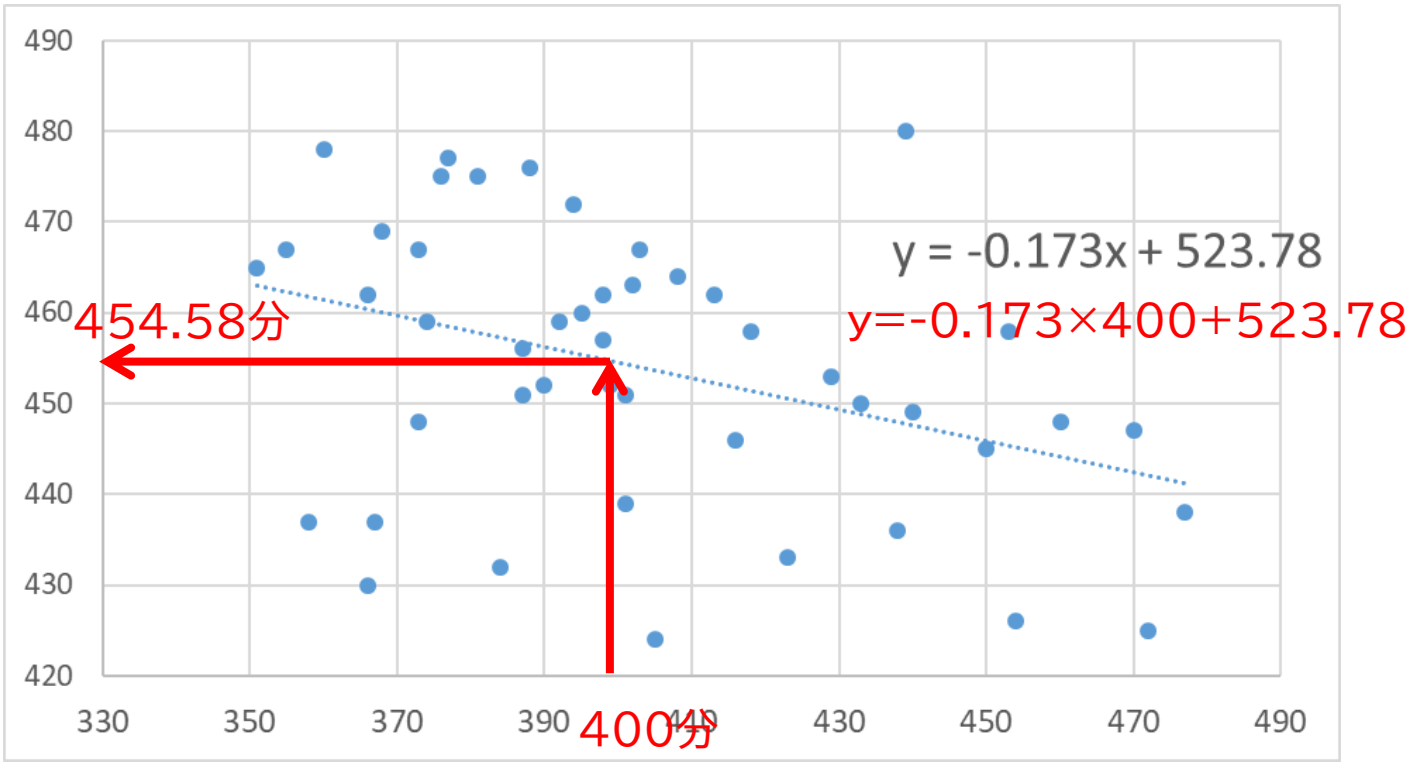
例）通学時間、買い物、休息・・・など
(補足：どの説明変数を利用するのか？(変数選択))

令和３年社会生活基本調査　生活時間－地域（調査票Ａ）
第６５－１表　曜日，男女，スマートフォン・パソコンなどの使用時間，年齢，行動の種類別総平均時間(10歳以上)－全国，都道府県

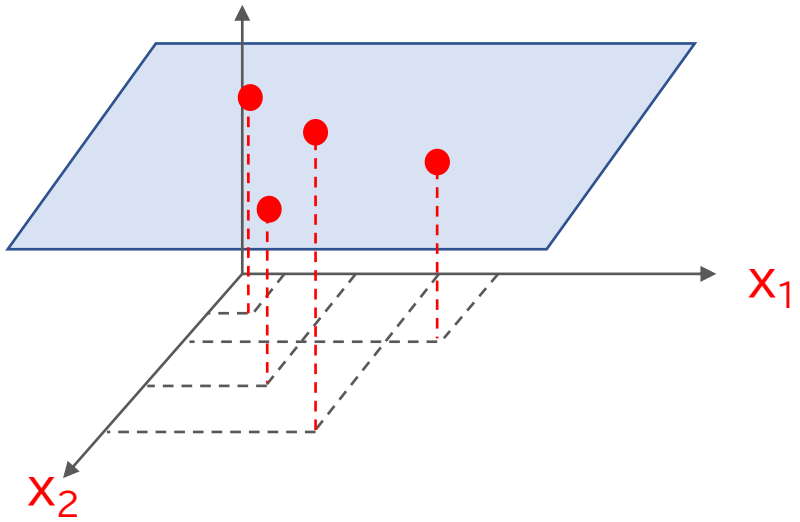
					総平均時間 行動の種類 01_睡眠 (分)	総平均時間 行動の種類 02_身の回りの用事 (分)	総平均時間 行動の種類 03_食事 (分)	総平均時間 行動の種類 04_通勤・通学 (分)	総平均時間 行動の種類 05_仕事 (分)	総平均時間 行動の種類 06_学業 (分)	総平均時間 行動の種類 07_家事 (分)
曜日	地域区分	男女	スマートフ	年齢							
2_平日	01_北海道	0_総数	0_総数	02_15～19歳	467	72	74	62	41	355	
2_平日	02_青森県	0_総数	0_総数	02_15～19歳	469	71	90	53	34	368	
2_平日	03_岩手県	0_総数	0_総数	02_15～19歳	458	69	76	53	47	453	
2_平日	04_宮城県	0_総数	0_総数	02_15～19歳	448	75	80	72	44	373	
2_平日	05_秋田県	0_総数	0_総数	02_15～19歳	467	61	84	45	29	373	
2_平日	06_山形県	0_総数	0_総数	02_15～19歳	477	74	83	52	43	377	
2_平日	07_福島県	0_総数	0_総数	02_15～19歳	436	60	87	64	41	438	
2_平日	08_茨城県	0_総数	0_総数	02_15～19歳	456	88	76	71	31	387	
2_平日	09_栃木県	0_総数	0_総数	02_15～19歳	472	66	82	69	46	394	
2_平日	10_群馬県	0_総数	0_総数	02_15～19歳	457	74	78	69	41	398	

2つの値を使って予測する

1つの値を使って予測



2つの値を使って予測



$$y=a_1x_1 + a_2x_2 + b$$

さらに多くの値を使って予測する

図として表すことはできないけれど
同じような考え方を使って予測をする

	1	2	3	4	5	6	7	8	9	10	11
1	地域区分	睡眠	身の回りの	食事	通勤・通学	仕事	学業	家事	買い物	移動(通勤	テレビ
2	01_北海道	467	72	74	62	41	355	5	3	19	
3	02_青森県	469	71	90	53	34	368	0	0	13	
4	03_岩手県	458	69	76	53	47	453	5	3	8	
5	04_宮城県	448	75	80	72	44	373	3	3	7	
6	05_秋田県	467	61	84	45	29	373	4	6	13	
7	06_山形県	477	74	83	52	43	377	4	5	6	
8	07_福島県	436	60	87	64	41	438	7	2	15	
9	08_茨城県	456	88	76	71	31	387	6	6	18	
10	09_栃木県	472	66	82	69	46	394	13	12	9	
11	10_群馬県	457	74	78	69	41	398	7	7	12	
12	11_埼玉県	439	83	91	80	31	401	5	3	12	

睡眠時間$=a_1\times$学習時間$+\ a_2\times$ 食事時間$+a_3\times$通学時間$+\cdots+\ b$

他にも例えば…

- アパートの家賃を推測しよう
  - 駅からの「距離」が関係しそう…：近いほど高い・遠いほど安い

    $(\text{家賃}) = a_1 \times \text{距離} + b$
  - 家賃は距離以外にも関係しそう…：築年数，間取り，コンビニのありなし，等

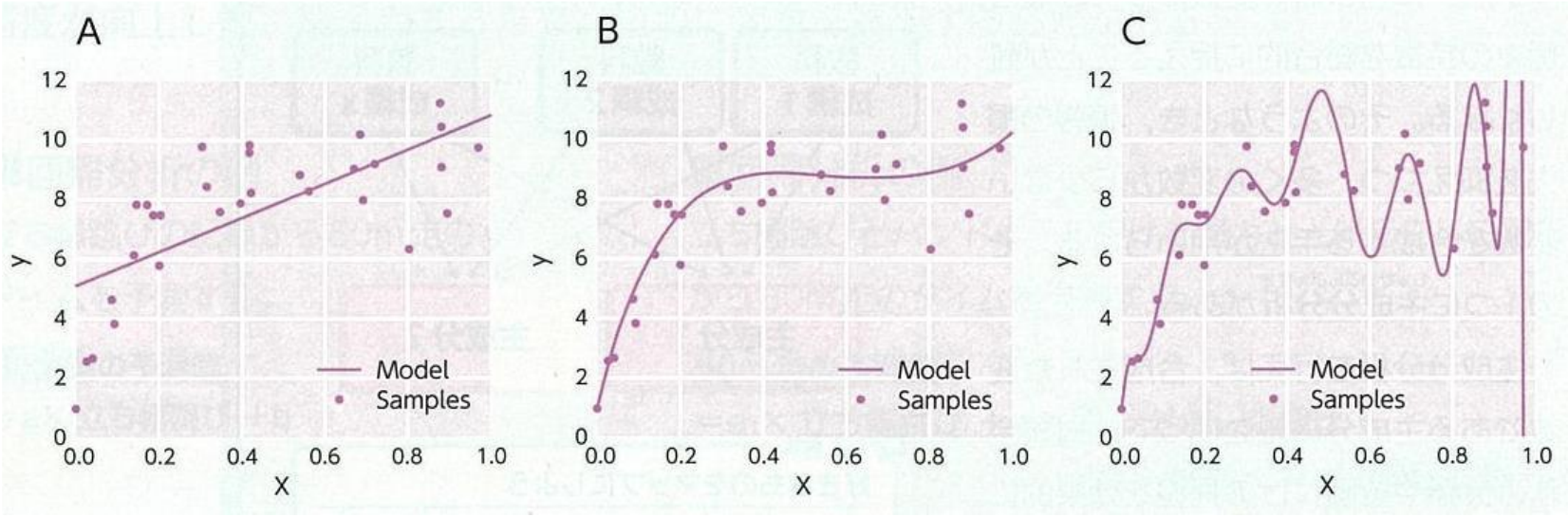
    $(\text{家賃}) = a_1 \times \text{距離} + a_2 \times \text{築年数} + b$

- 自分の街の魅力度を推測しよう
  - 魅力度を決定するような要因(変数)は何だろうか？
    - 人口が多いこと？
    - 公園や自然が多いこと？
    - ショッピングセンターがあること？
    - 公的施設が充実していること？
  - いろいろな変数がある中で，どれを採用すればよいかと考えると変数選択も分析可能

回帰がうまくいかない時

- 回帰分析を実施してもうまくいかないケースもある
- 当てはまりが良くない(決定係数R^2の値が良くない)

→ 変数を別のものに変える？変数の数を増やすとよいかも？
そもそも線形では表現できない？



東京書籍「情報II」
84ページより抜粋

- 当てはまり良すぎるのも問題(**過学習**)

学習時のデータに対してはよい精度を出すが、未知データに対しては同様の精度を出せない

回帰がうまくいかない時

- 回帰分析を実施してもうまくいかないケースもある
- 当てはまりが良くない(決定係数R^2の値が良くない)

  → 変数を別のものに変える？変数の数を増やすとよいかも？
  　そもそも線形では表現できない？

- 変数を増やしていくと**多重共線性**の心配が…

  （例） 睡眠時間＝a_1×学習時間＋ a_2× 食事時間
  　　＋ a_3×通学時間 ＋ a_4×公共交通機関の乗車時間 ＋ … ＋ b

- もし「通学時間」と「公共交通機関の乗車時間」が両方含まれていたら…

  → この2つは相関がとても高い(「通学時間」≒「公共交通機関の乗車時間」)

  → それぞれの係数の値(a_3とa_4の値)が不安定になる(参考：VIFを確認)

結果が良くない時こそ，分析を深堀するビッグチャンス！

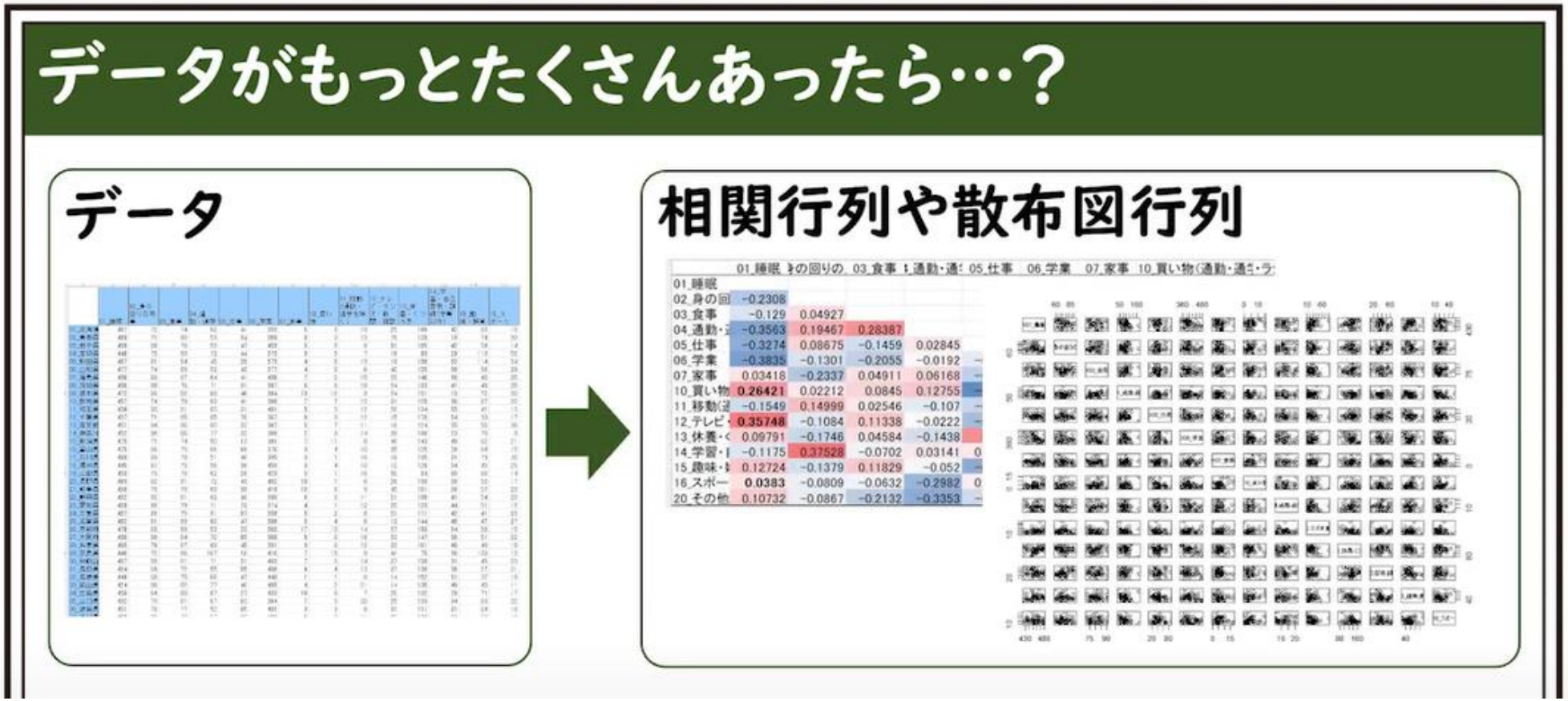
情報とデータサイエンス

主成分分析による次元縮約

データに関しては統計センターの
「SSDSE（教育用標準データセット）」家計消費を利用
https://www.nstac.go.jp/use/literacy/ssdse/#SSDSE-C

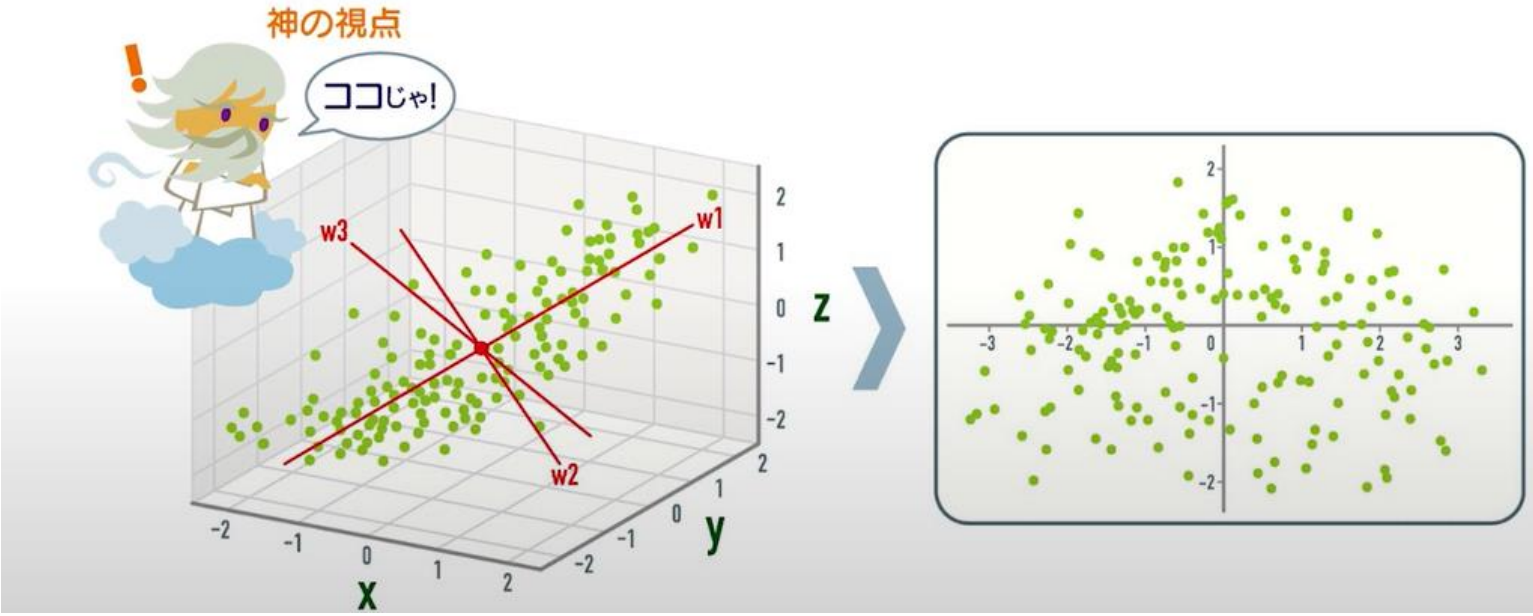
データの特徴を説明したい

- 2変数間であれば相関を見ればある程度把握できるが，全体の特徴を知るにはどうすればよい？



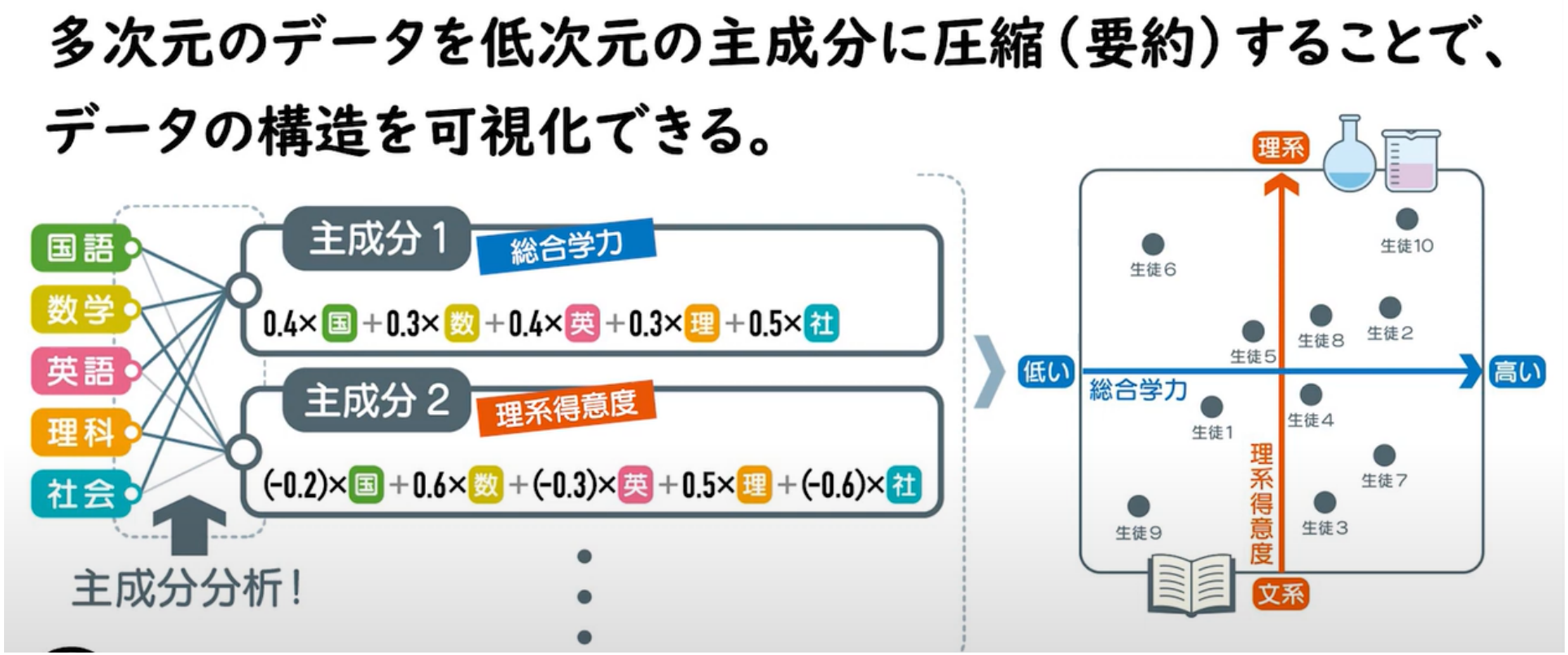
たくさんの変数をうまく圧縮したい

- 国語，数学，英語，理科，社会の成績から，学生の特徴を分析したい
- 説明変数が5つ(=5次元)なので，図示ができない
  → 変数をうまく圧縮(合成・統合)して，図示することで分析を進めたい
  → 圧縮の方法はたくさんあるが，どの方法がベター？
  → データが最も散らばるような変数の統合方法(**主成分分析**)



第1主成分(w1)：分散が最も大きい(バラツキが最も大きい)軸
第2主成分(w2)：分散が2番目に大きい軸

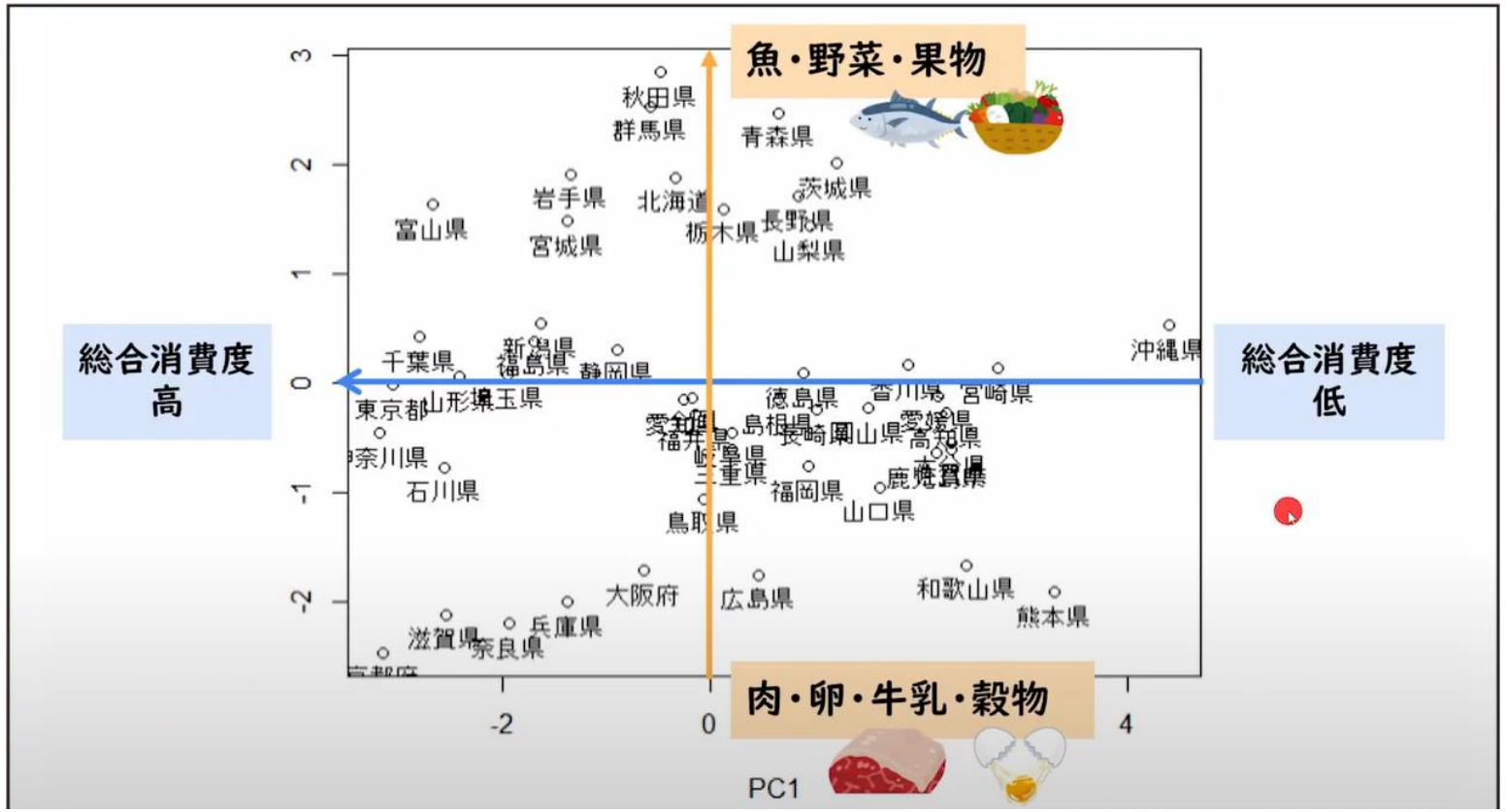
主成分分析のポイント



- それぞれの主成分が何を表しているかは人間が行う(係数の値を見ながら)
- 主成分の寄与率を見ることで，実際のデータのどのくらいの情報を表現できているかが分かる

他の例でも確認してみよう

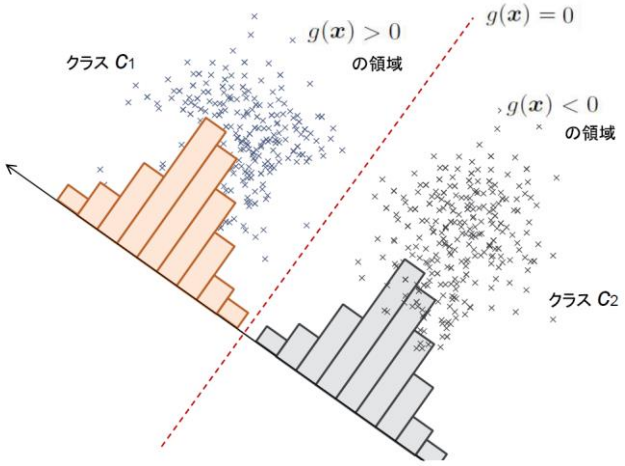
- SSDSE（教育用標準データセット）」家計消費を利用した分析
  (詳細は講義動画参照)



(参考)次元圧縮の方法はさまざま

- 次元圧縮の方法(統計的手法)
  - 主成分分析(Principal Component Analysis，PCA)
    元の情報をできるだけ失わないように，本質にかかわる成分を抽出する次元削減法

  - 線形判別分析(Linear Discriminant Analysis，LDA)
    次元削減をするが、クラス間の分散を最大にし
    クラス内の分散を最小にする軸を使うことで
    分類を容易にする次元削減法



  - 潜在意味解析(Latent Semantic Analysis，LSA)
    行列の特異値分解を利用して，複数の要素から成る成分を，
    特徴を捉えた小さな要素に分解する

情報とデータサイエンス

クラスタリング

データにクラスラベルがついていない教師なしデータにおけるクラスタリングの説明

データの特徴をつかむ別の方法

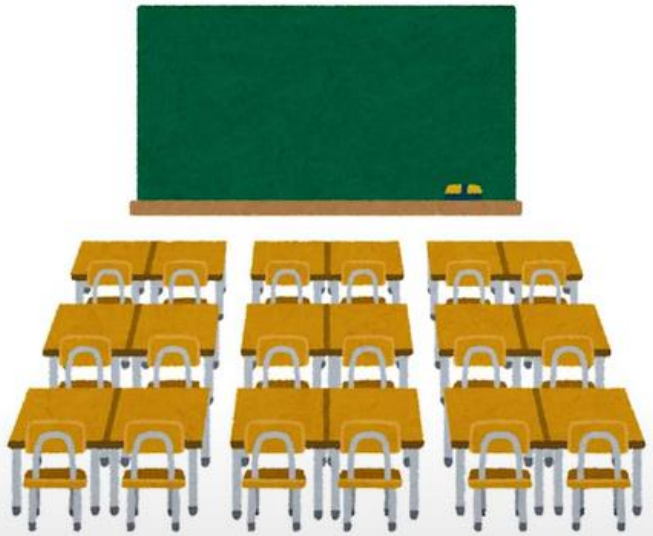
- 特徴が似たようなデータどうしをグループに分類した方が分析しやすい
- グループラベルが無い場合，特徴が似ている＝データ間の距離が近い

  (距離やデータの類似度でよく用いられるもの)
  - ユークリッド距離
  - マンハッタン距離
  - コサイン類似度
  - Jaccard係数(集合と集合の類似度)
  - KLダイバージェンス(確率分布間の類似度)

- グループを作る(＝クラスタリング)方法も様々
  - 階層的クラスタリング：デンドログラム
  - 非階層クラスタリング：k-means法(クラスタの平均を用い、与えられたクラスタ数k個に分類する)

クラスタリングの実例

クラスのみんな、気が合う人は誰だろう？



アンケートをやってみよう！

どれくらい好きか、選んで下さい。

5（とても好き） / 4(好き） / 3（普通）
2(あまり好きではない） / 1（好きではない）

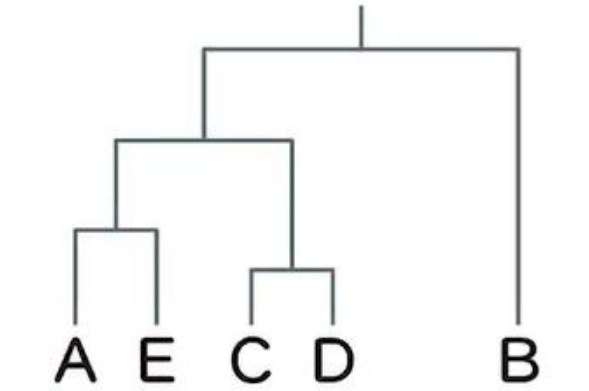
Q1. プログラミング　5 4 3 2 1

Q2. 野球　5 4 3 2 1

Q3. 楽器演奏　5 4 3 2 1

・
・
・

	洋楽	ドラマ	野球	JPOP	プログラミ	ライブ	バスケ	カラオケ	ペンギン	水泳	登山	舞台鑑賞	パン
生徒1	3	5	5	2	1	4	1	2	3	3	3	2	
生徒2	3	3	3	2	3	1	5	1	2	5	3	5	
生徒3	1	5	4	3	2	3	4	4	1	4	2	4	
生徒4	4	3	5	2	1	3	5	4	1	2	2	2	
生徒5	2	4	2	3	4	1	3	4	3	4	5	4	
生徒6	3	5	1	2	5	2	5	5	2	1	3	5	
生徒7	4	2	1	1	2	5	5	4	3	3	5	2	
生徒8	2	3	3	3	1	3	1	3	4	3	4	5	

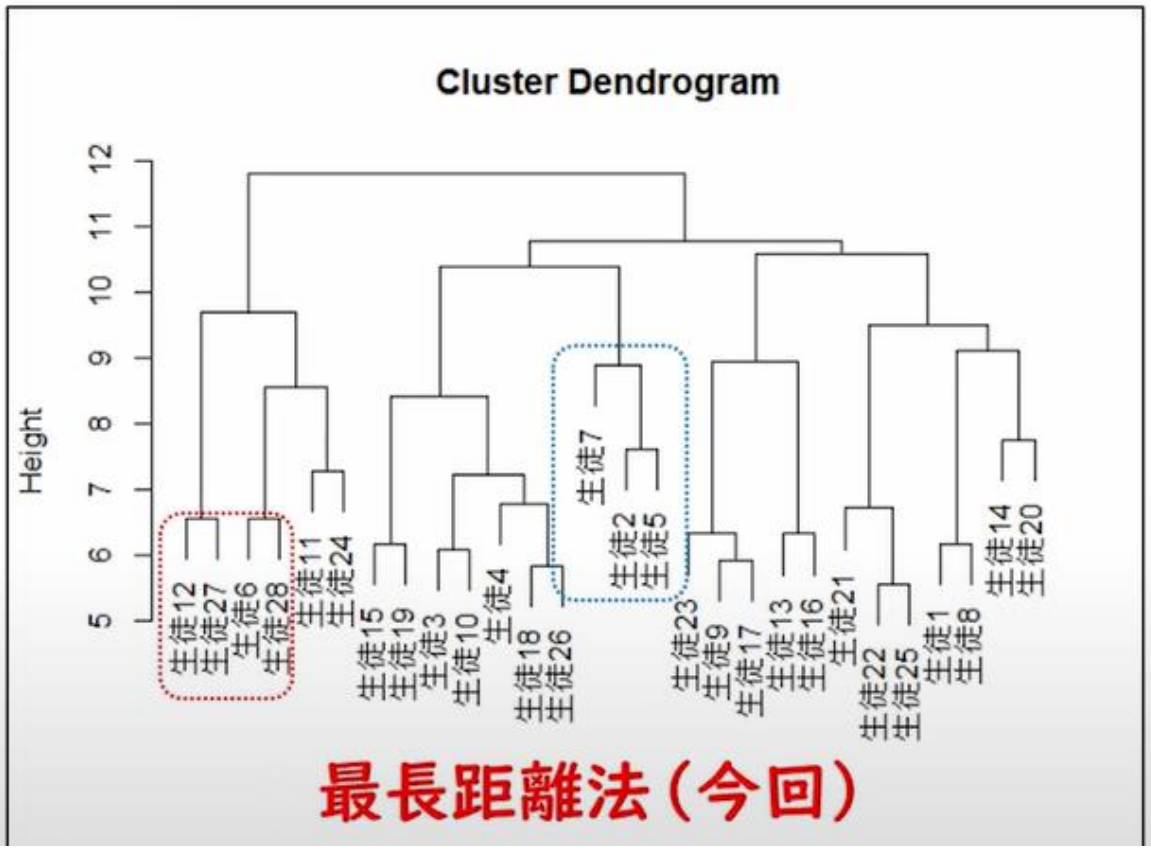


デンドログラム（樹形図）

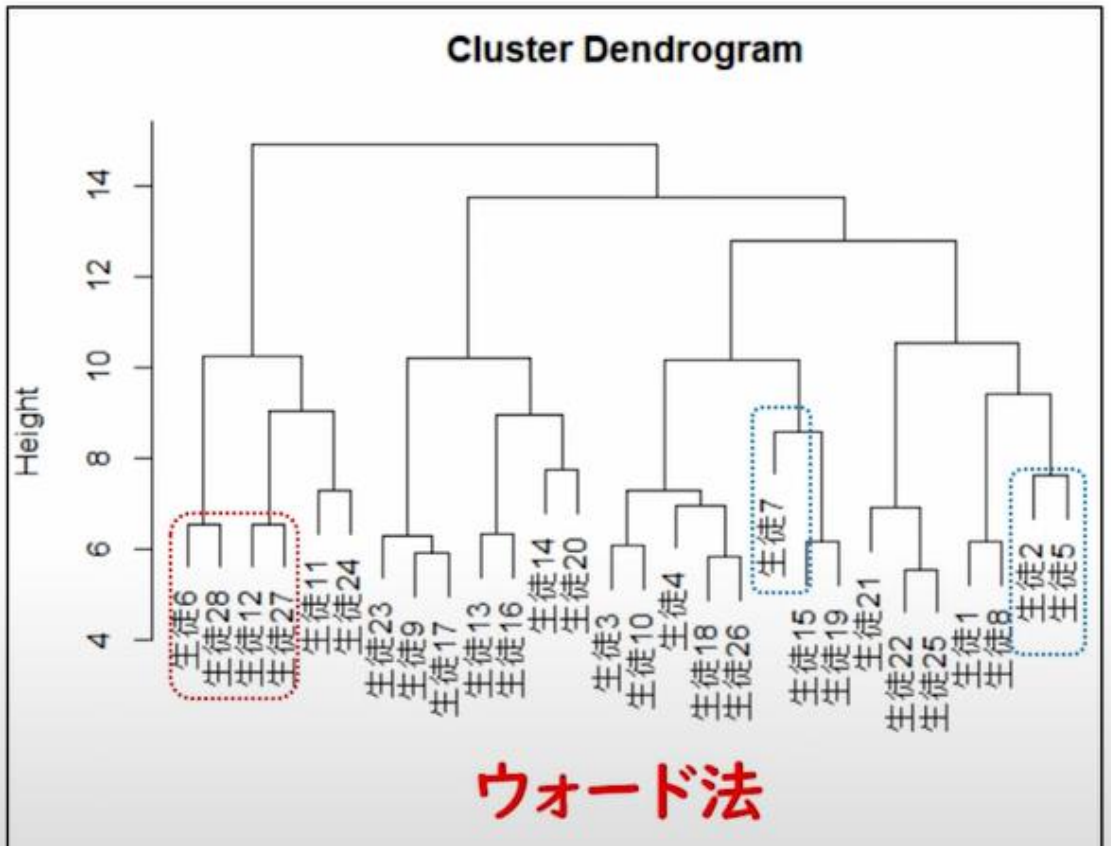
クラスタリングの実例

クラスタリングには様々な手法がある。

hclust(○○○○,method=“ward.D2”)

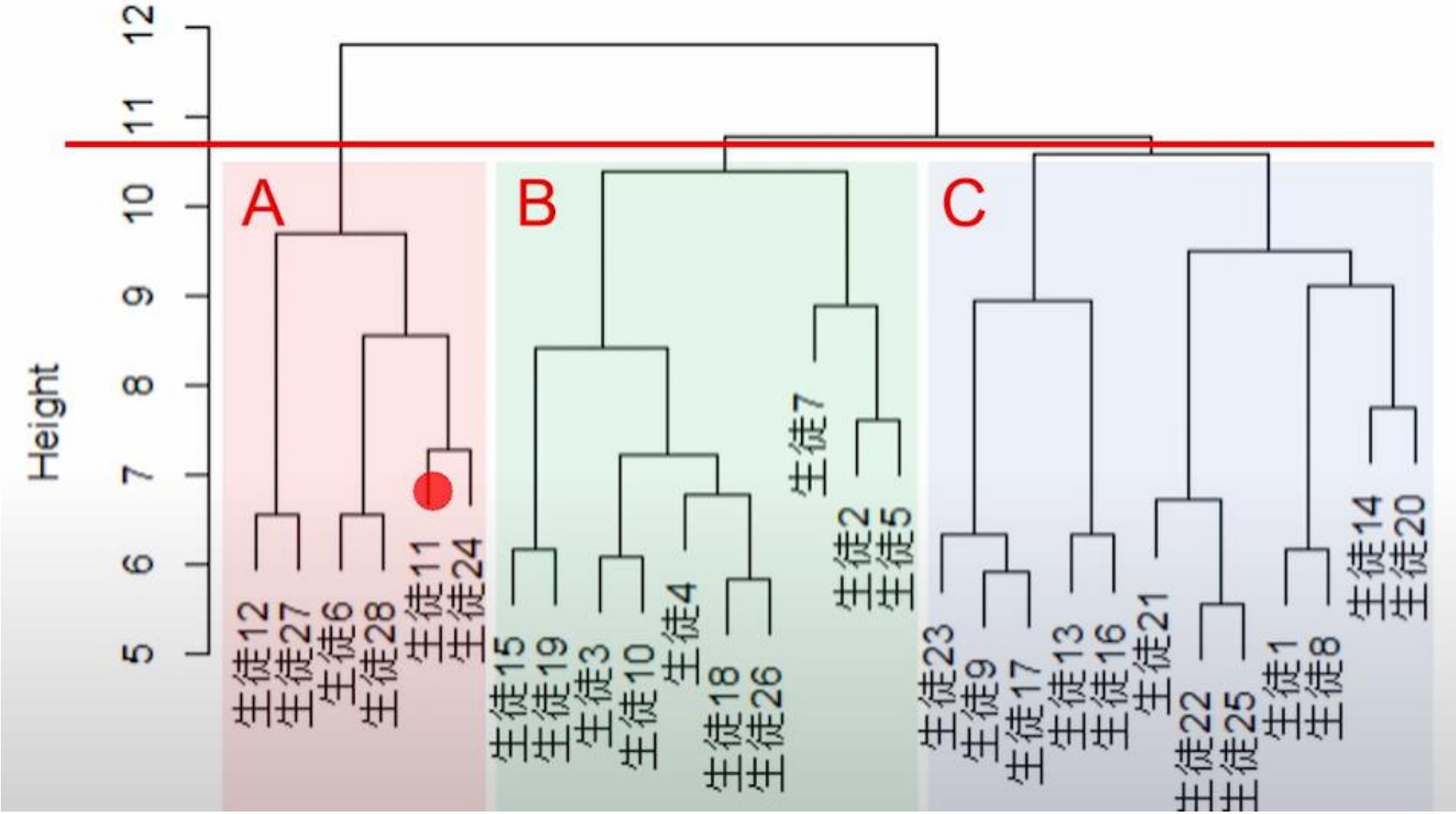


最長距離法（今回）



ウォード法

クラスタリングの実例


Height
12
11
10
9
8
7
6
5
A
B
C
生徒12
生徒27
生徒6
生徒28
生徒11
生徒24
生徒15
生徒19
生徒3
生徒10
生徒4
生徒18
生徒26
生徒7
生徒2
生徒5
生徒23
生徒9
生徒17
生徒13
生徒16
生徒21
生徒22
生徒25
生徒1
生徒8
生徒14
生徒20

クラスタリングを行ったあとが重要

- クラスタリングはあくまでグループ分け
  → グループ内・クループ間の特徴や差異を分析することが重要

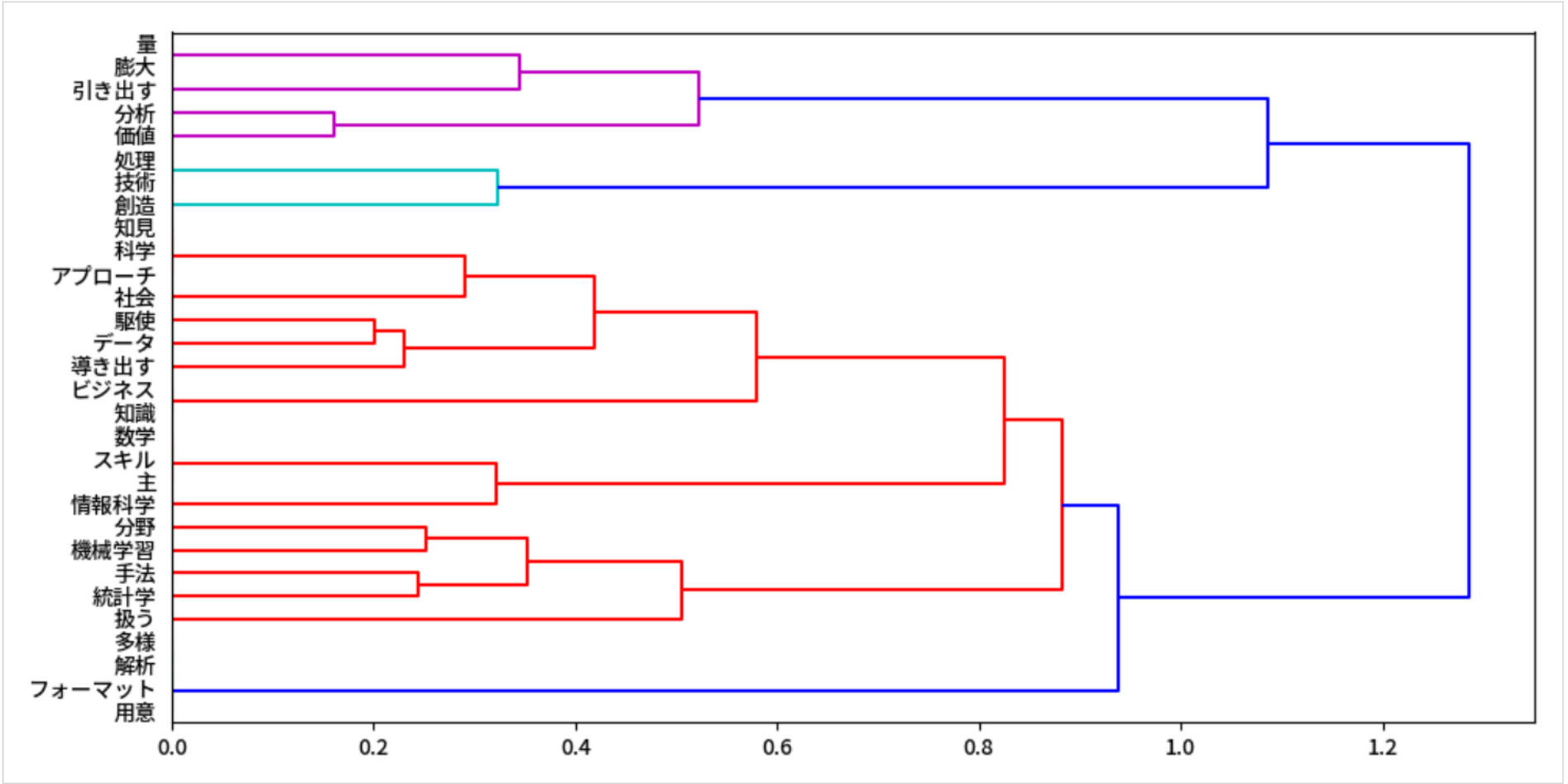
	洋楽	ドラ	野球	JPOP	プロ	ライ	バス	カラ	ペン	水泳	登山	舞台	パン	ダン	料理	楽器	昼寝	動画	犬	猫
a	3.6	2.2	3.5	2.4	2.3	1.9	2.3	3.1	3	3.4	3	2.6	3.2	3.1	4.5	2.8	2.3	3.2	3.3	3.1
b	2.8	3.5	2.8	2.3	1.8	2.8	4.5	3.8	2.1	3.5	3.4	3	2.8	3.7	2.8	2.6	3.6	3.3	3.2	2.2
c	4.2	3.5	1.8	4.5	3.2	3.2	2.3	4.5	2.7	2	2	4	2.7	4	3	4.7	3.7	4	3.3	3.3

- デンドログラムにおける境界線(閾値)の与え方，k-meansにおけるkの与え方次第でグループ数も分けられるグループも変化する

ちなみに最初に示しましたが…

階層的クラスタリング(β版)

文章中での出現傾向が似た単語をまとまりとしてとらえられるよう樹形図で表したものです。グループは色分けして表示しています。



(出典)　https://textmining1.userlocal.jp/home/result/129a35b46ac0638ece0c62f2c64bbc2a

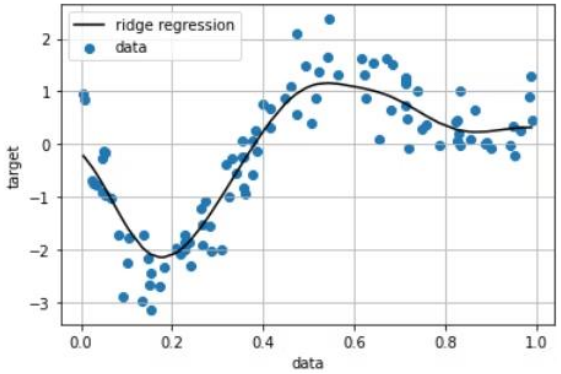
統計的手法の現在の潮流

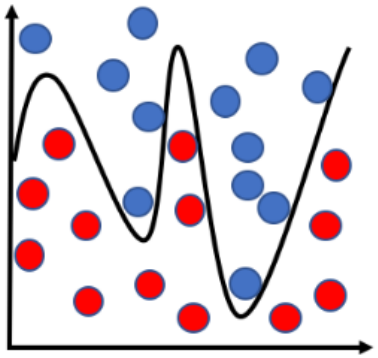
- 目的変数がある場合 → 予測や分類

		目的変数	
		量的変数	質的変数
説明変数	量的変数	単回帰分析 **重回帰分析**	**判別分析** ロジスティック回帰
	質的変数	**数量化I類**	**数量化II類**
		⇩	⇩
		回帰問題	**分類問題**

- ここまでの話は，線形モデルの枠組みなので，シンプルなモデル

⇨ **より複雑なモデル，非線形のモデルも扱いたい**





統計的手法の現在の潮流

- 目的変数がある場合 → 予測や分類

		目的変数	
		量的変数	質的変数
説明変数	量的変数	単回帰分析 **重回帰分析**	**判別分析** ロジスティック回帰
	質的変数	**数量化I類**	**数量化II類**
		⇩	⇩
		回帰問題	**分類問題**

- ここまでの話は，線形モデルの枠組みなので，シンプルなモデル

⇨ **より複雑なモデル，非線形のモデルも扱いたい** ⇨

一般化線形モデル
階層ベイズモデル
混合回帰モデル
一般化加法モデル(GAM)
GA2M (Generalized Additive 2 Model)
Factorization Machine　など

(参考)検定

Q：このサイコロって本当に偏りのないサイコロ？

- 120回振って，出た回数を集計

目	1	2	3	4	5	6
回数	10	11	26	15	28	30
理想	20	20	20	20	20	20

- 偏りがないかどうか(理想的かどうか)を統計的に検定(今回は適合度検定)

- 検定のポイントは…
  - どんな仮説を検定したいのか？(今回は偏りがあるかどうか)
  - 母集団がどのような分布に従っているのか？
  - 検定で用いる指標(検定統計量)がどのような確率分布に従っているのか？(今回の適合度検定は，χ^2乗分布に従う)

本日のまとめ

- 「**統計的手法をデータ分析(何が重要？どれが効いてる？)に活かす**」
  - 予測したい・目的(結果)に対して何が効いてるか知りたい　→　**回帰分析**
  - 変数を減らして(統合して)うまく説明したい　→　**主成分分析**
  - データをうまく分類したい　→　**クラスタリング**
  - 仮説が正しいかチェックしたい　→　**検定(推定)**

- Web上で手法の解説やプログラム等を掲載してくれているため，手法を「使う」面である程度のプログラミング力があれば，特に難しくない(はず)
- 「なぜその手法を使うのか」を理解できることも，今後データサイエンスに関わっていくとすれば重要

まずは今回紹介(講義動画で紹介)の方法からチャレンジしてみましょう！