

情報科学 【AI・データサイエンス】

第6回 データの可視化

可視化とは？
いろいろな可視化手法

可視化とは？

数字だけ並べてもよくわからない → 絵にするとよくわかる！

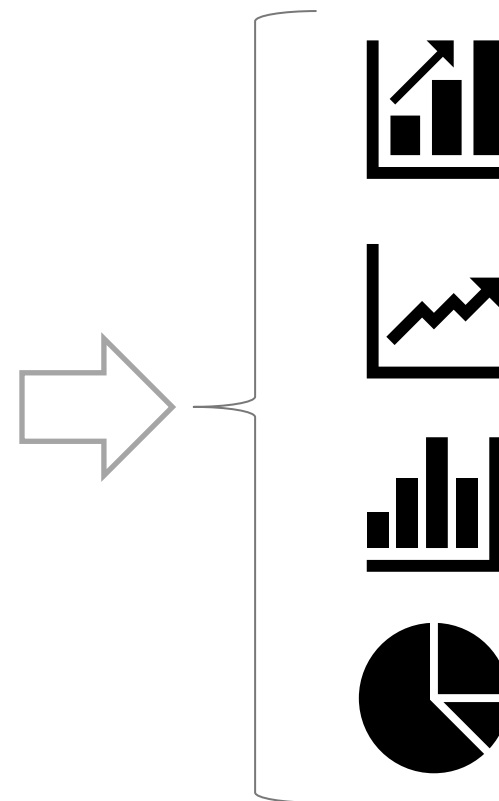
(再び) データとは

- 定義（デジタル広辞苑より）
 1. 物事の推論の基礎となる事実 また、参考となる資料・情報
 2. コンピューターで、
プログラムを使った処理の対象となる記号化・数字化された資料
- 例
 - 身長と体重
 - 情報科学のテストの点数
 - 通販サイトの購入・閲覧履歴
 - 授業アンケートの回答
 - 授業の感想文
 - SNSに投稿した写真
 - このページを見ている皆さんの教科書閲覧時間

可視化

- データを直感的に理解できる図にすること
 - あらゆる分野で必須技術

ID	属性1	属性2	属性3	属性4	...	属性N
0	0.60	0.39	0.54	0.79	...	0.34
1	0.15	0.14	0.04	0.01	...	0.23
2	0.42	0.71	0.62	0.39	...	0.66
3	0.35	0.59	0.30	0.12	...	0.31
4	0.05	0.23	0.53	0.33	...	0.82
5	0.88	0.16	0.35	0.48	...	0.91
6	0.40	0.80	0.84	0.65	...	0.94
7	0.59	0.73	0.04	0.34	...	0.22
8	0.61	0.56	0.06	0.52	...	0.00
9	0.55	0.70	0.70	0.64	...	0.32
⋮	⋮	⋮	⋮	⋮	⋮	⋮



体重と身長の測定データ

氏名	体重 (kg)	身長 (cm)	氏名	体重 (kg)	身長 (cm)
A	31	127.4	N	29	138.7
B	34.7	137.3	O	20.9	139.4
C	54.7	155.6	P	27.9	138.6
D	24.3	127.1	Q	44.9	147.5
E	26.9	132.5	R	47.8	166.3
F	30.9	137.5	S	52.3	170.5
G	35.6	140.2	T	60.2	176.1
H	59.8	162.7	U	54	178.6
I	55.8	162.3	V	19.2	114.8
J	19.1	110.8	W	15	104.3
K	18.3	120.9	X	20.9	107.9
L	18.4	118	Y	58.6	190.5
M	62.5	166.8	Z	66.2	166.8



可視化の必要性

- 膨大なデータの把握は困難
 - 事例の多さ
 - たった26人の身体測定でも大変！
 - 属性の多さ
 - 身体測定では、多くの測定項目がある（例えば、性別、年齢、血圧、…）

氏名	体重 (kg)	身長 (cm)	氏名	体重 (kg)	身長 (cm)
A	31	127.4	N	29	138.7
B	34.7	137.3	O	20.9	139.4
C	54.7	155.6	P	27.9	138.6
D	24.3	127.1	Q	44.9	147.5
E	26.9	132.5	R	47.8	166.3
F	30.9	137.5	S	52.3	170.5
G	35.6	140.2	T	60.2	176.1
H	59.8	162.7	U	54	178.6
I	55.8	162.3	V	19.2	114.8
J	19.1	110.8	W	15	104.3
K	18.3	120.9	X	20.9	107.9
L	18.4	118	Y	58.6	190.5
M	62.5	166.8	Z	66.2	166.8



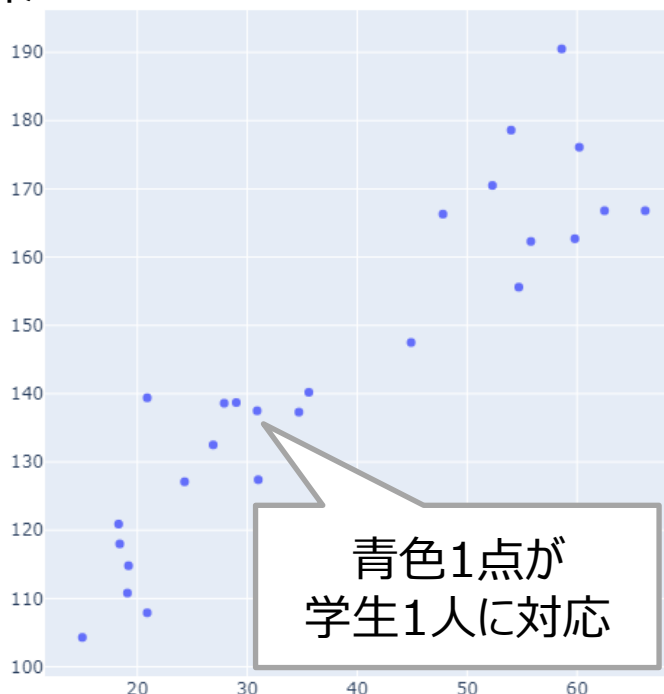
可視化によるデータ傾向の把握

- 身長と体重を**散布図**で描く

氏名	体重 (kg)	身長 (cm)	氏名	体重 (kg)	身長 (cm)
A	31	127.4	N	29	138.7
B	34.7	137.3	O	20.9	139.4
C	54.7	155.6	P	27.9	138.6
D	24.3	127.1	Q	44.9	147.5
E	26.9	132.5	R	47.8	166.3
F	30.9	137.5	S	52.3	170.5
G	35.6	140.2	T	60.2	176.1
H	59.8	162.7	U	54	178.6
I	55.8	162.3	V	19.2	114.8
J	19.1	110.8	W	15	104.3
K	18.3	120.9	X	20.9	114.8
L	18.4	118	Y	58.6	166.3
M	62.5	166.8	Z	66.2	176.1



第2要素
(身長)



右肩上がりの傾向に見える！

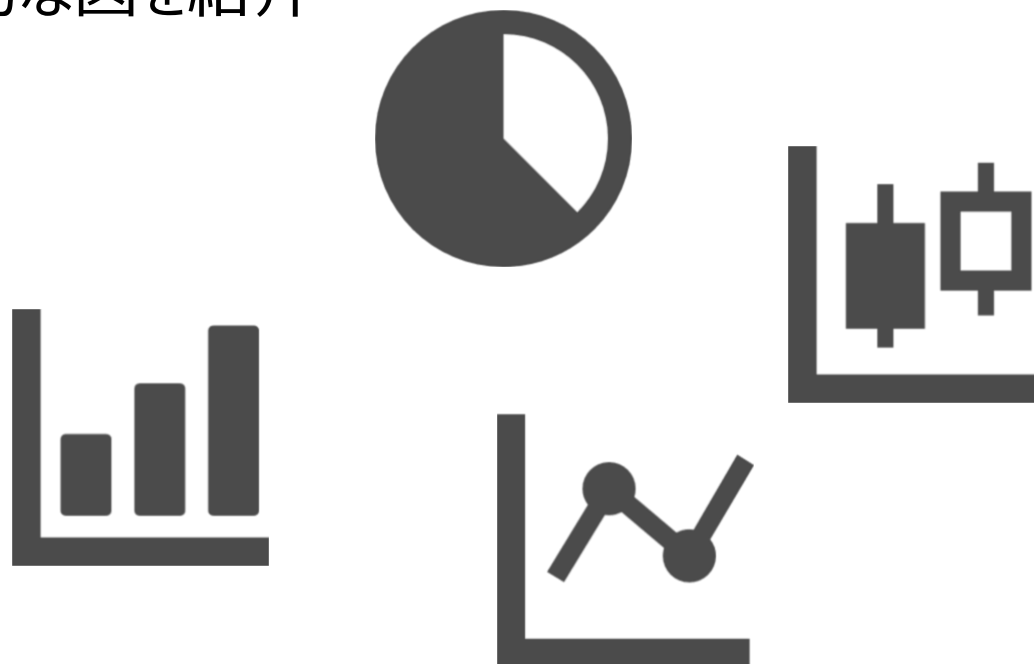


第1要素
(体重)

可視化手法

- 本講義では代表的な図を紹介

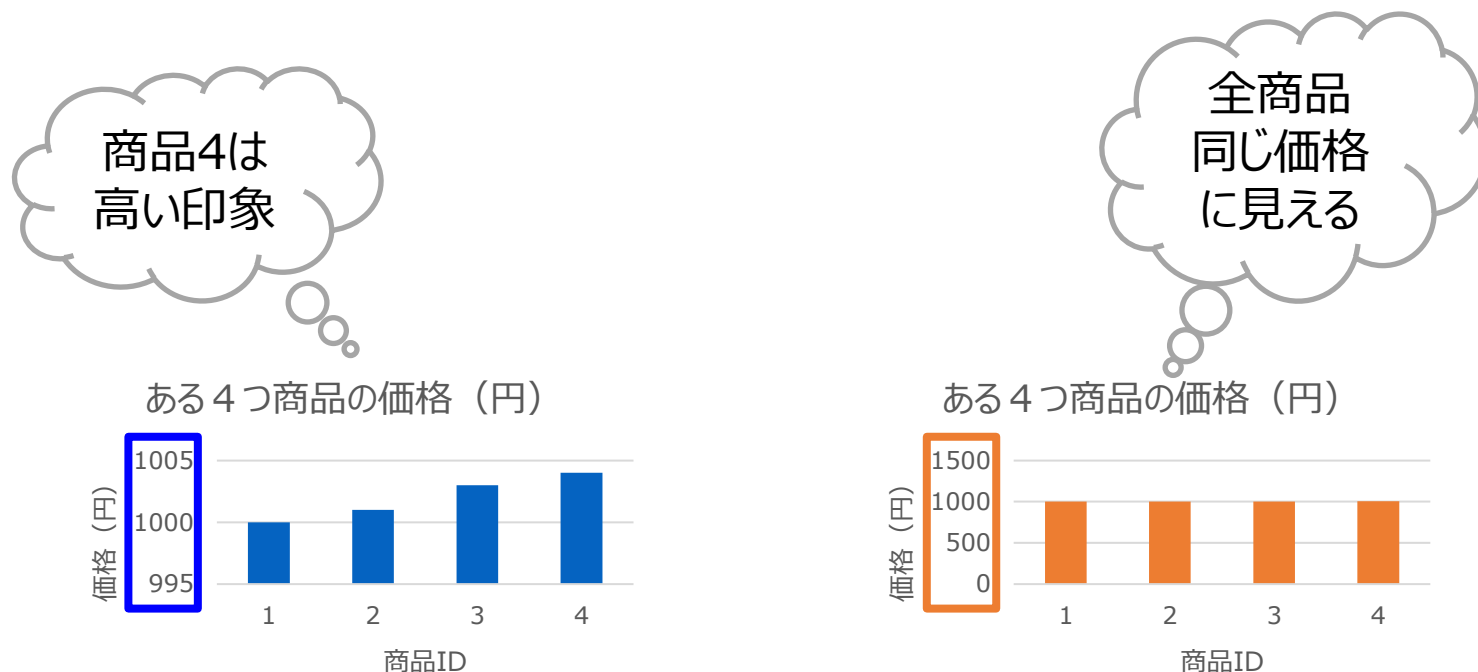
- ヒストグラム
- 箱ひげ図
- 棒グラフ
- パイチャート
- 散布図
- ヒートマップ
- 折れ線グラフ
- 無向/有向グラフ



＊以降の例で身長と体重データは疑似的に生成したデータ

可視化における注意点： 目的と条件によって適切なものを選択するべし

- 不適切な可視化は、誤解を生んだり、不誠実な印象を招く
- 各々の可視化手法がどのような効果をもち、どのような場合には使うか把握する



同じデータなのに縦軸の表示の違いで印象が違う！？

いろいろな可視化手法

またもや「ベクトル」登場

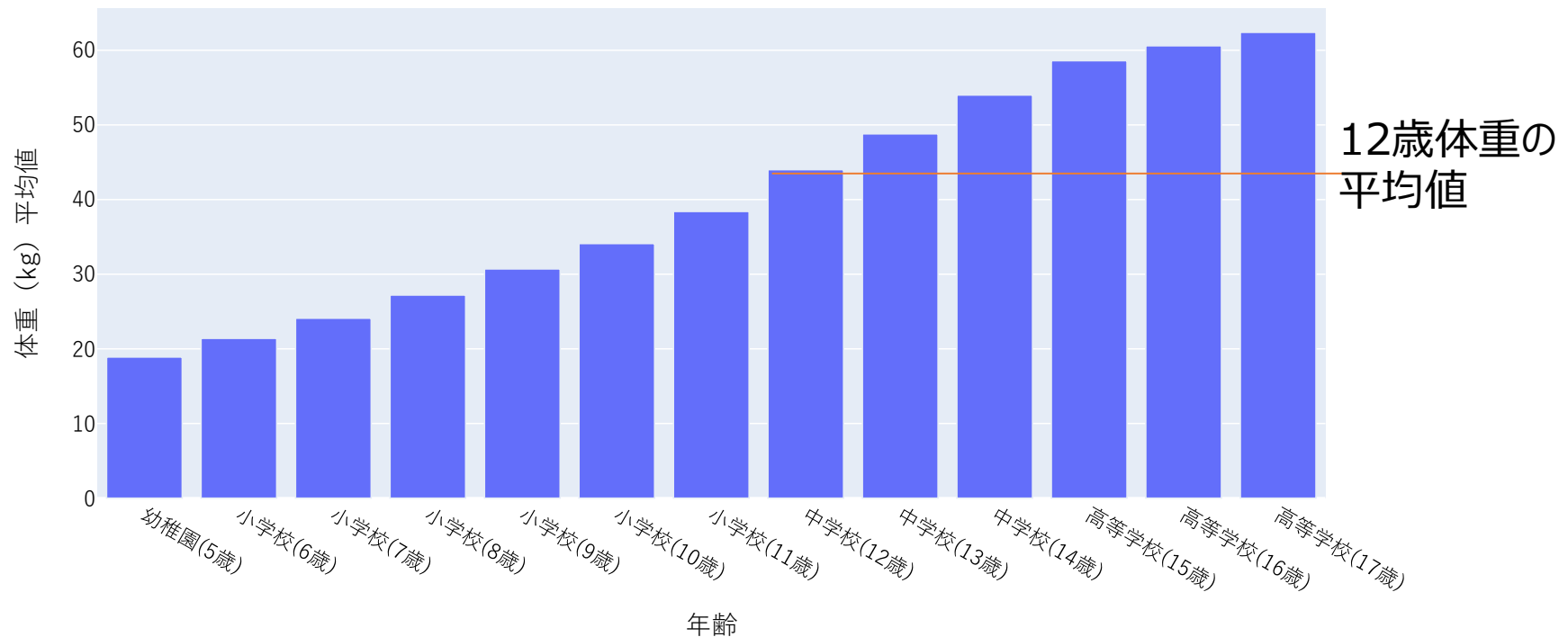
棒グラフ

最も基本

棒グラフ

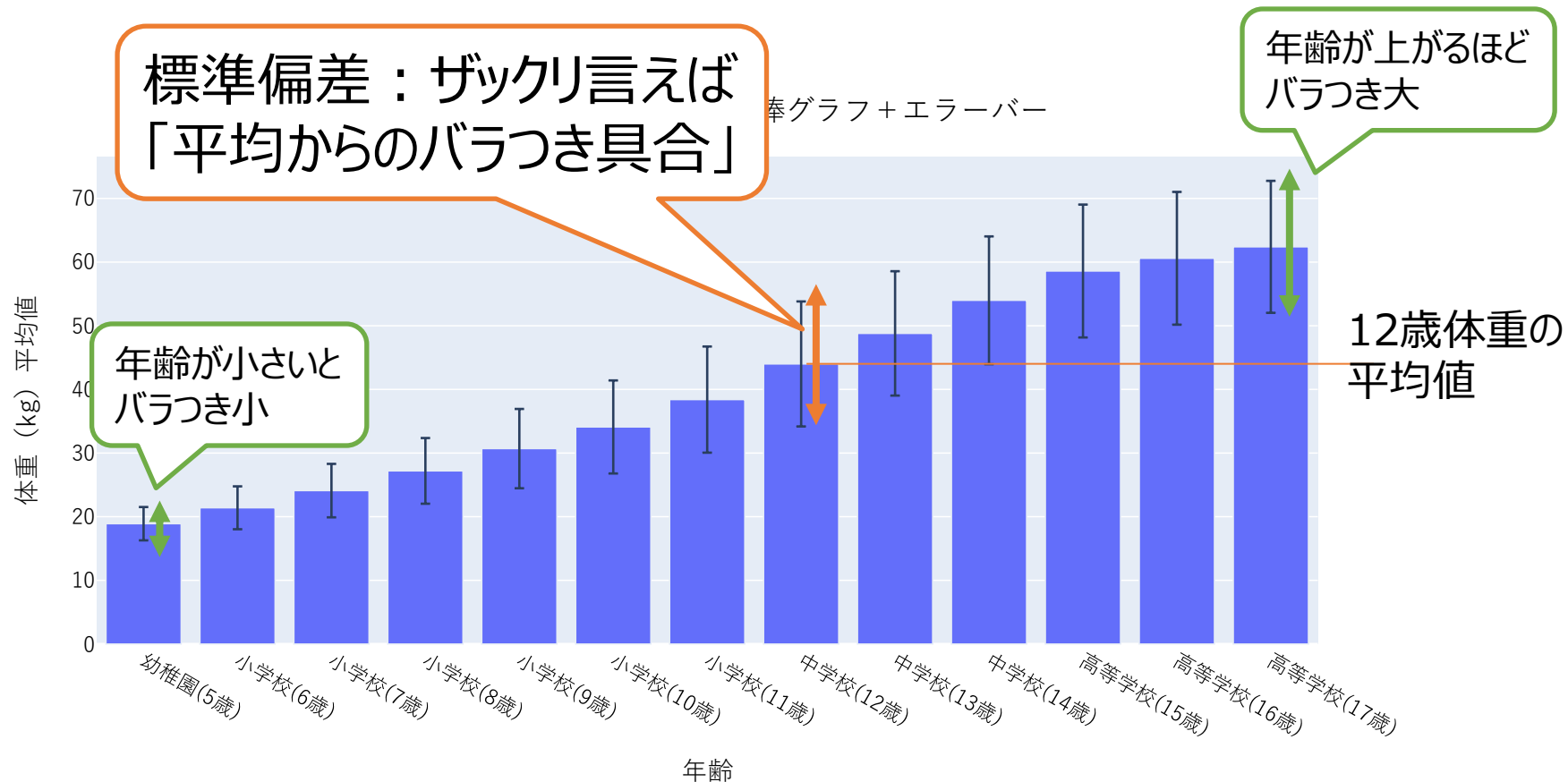
- 棒の高さで値を表現した図
- 数値データの比較を行う場合に有効
 - 平均体重は年齢でどのくらい違うのか見たい！

2018年度 男性 体重の年齢別 棒グラフ

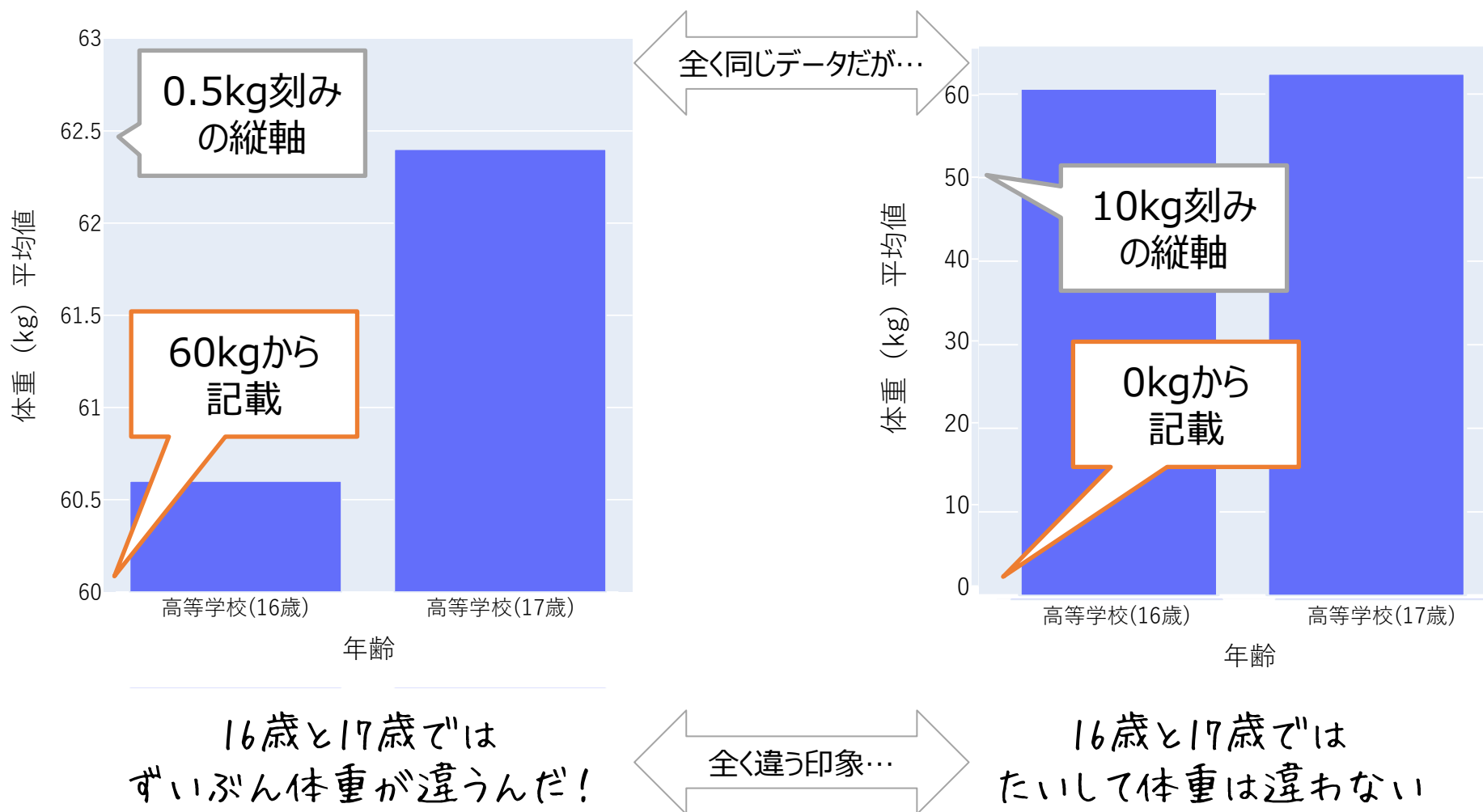


棒グラフ+エラーバー

- **標準偏差**でデータの散らばり具合を表現できる
 - 平均値と合わせて利用されることが多い



棒グラフの縦軸に関する注意： 縦軸の原点を恣意的に選ぶべきではない



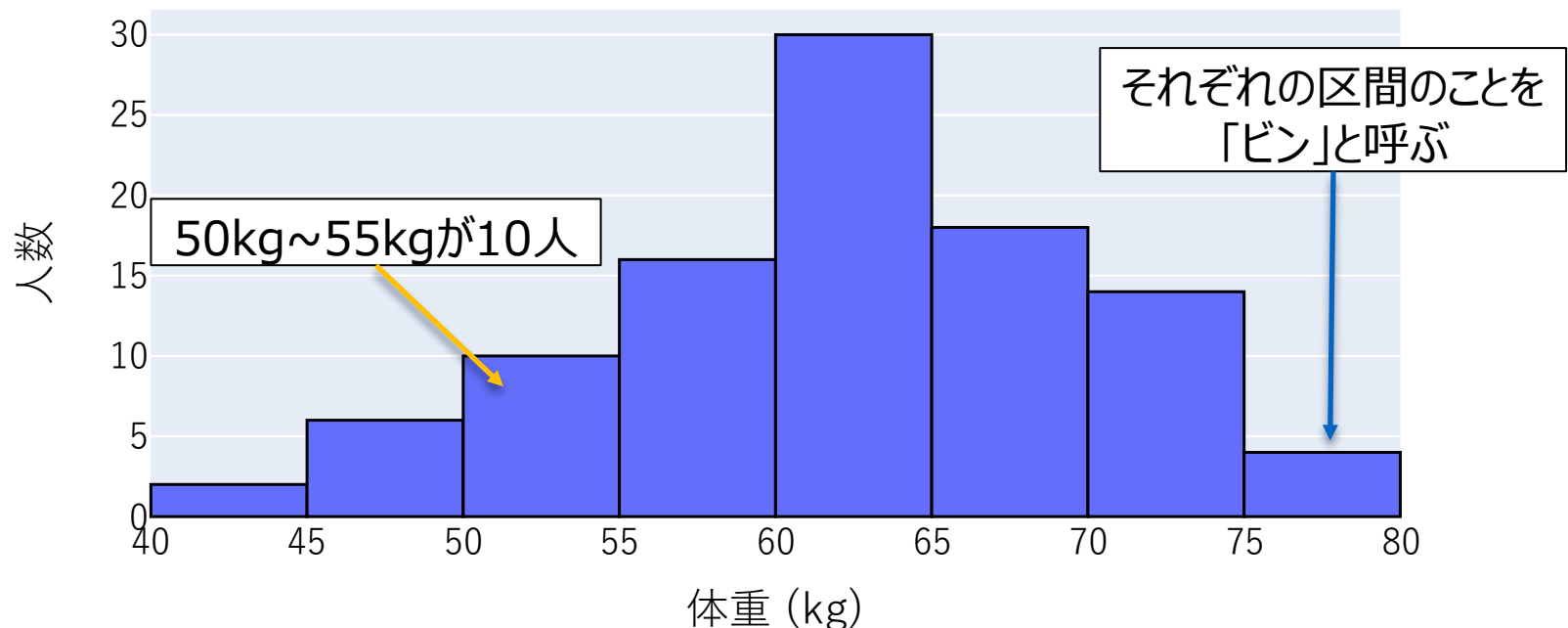
ヒストグラム

「頻度・回数」を表す棒グラフ

ヒストグラム＝「頻度（回数）」を表す棒グラフ

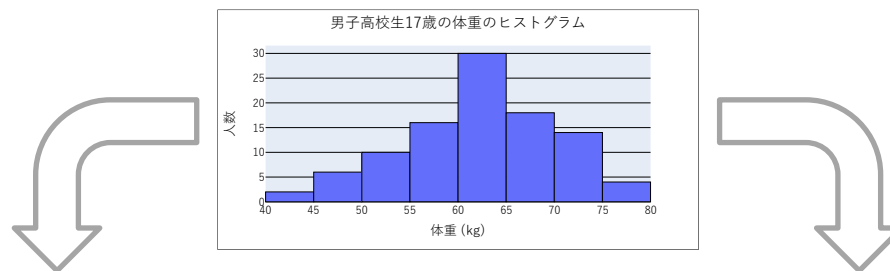
- 区間ごとにデータが観測された回数を集計した図
- どんな値がどのくらいあるのか＝データの分布を調べる場合に有効
 - どんな体重の人が多いのか知りたい！

男子高校生17歳の体重のヒストグラム

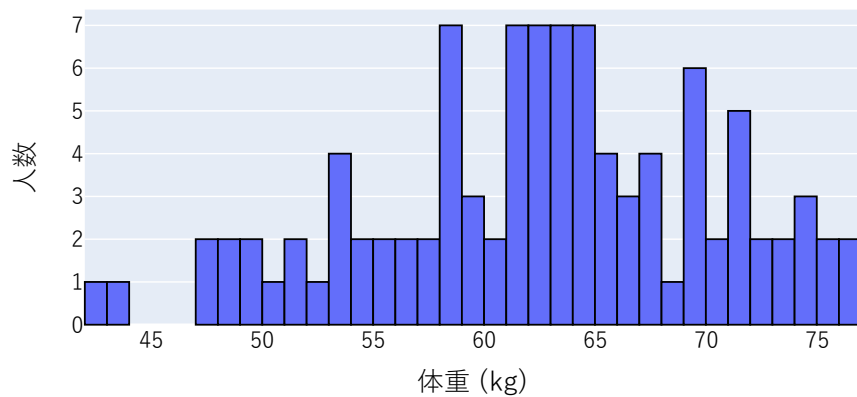


ビン幅によって見た目が大きく変わることに注意

- 経験的にデータ数の平方根程度が一つの目安

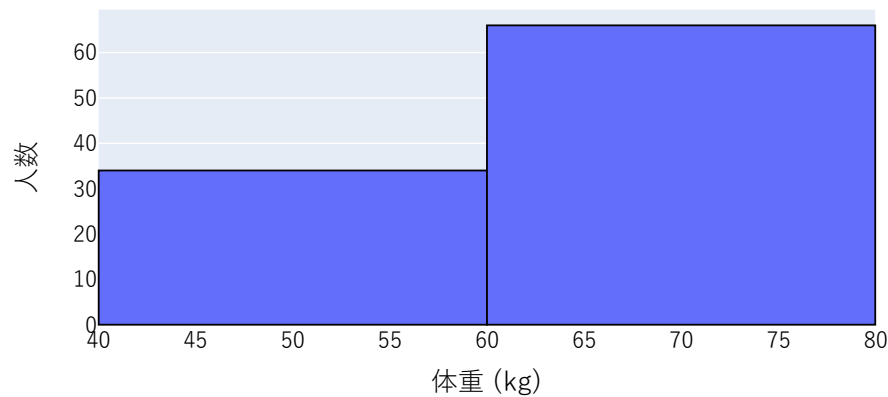


男子高校生17歳の体重のヒストグラム（ビン幅狭い）



傾向を把握しにくい

男子高校生17歳の体重のヒストグラム（ビン幅広い）

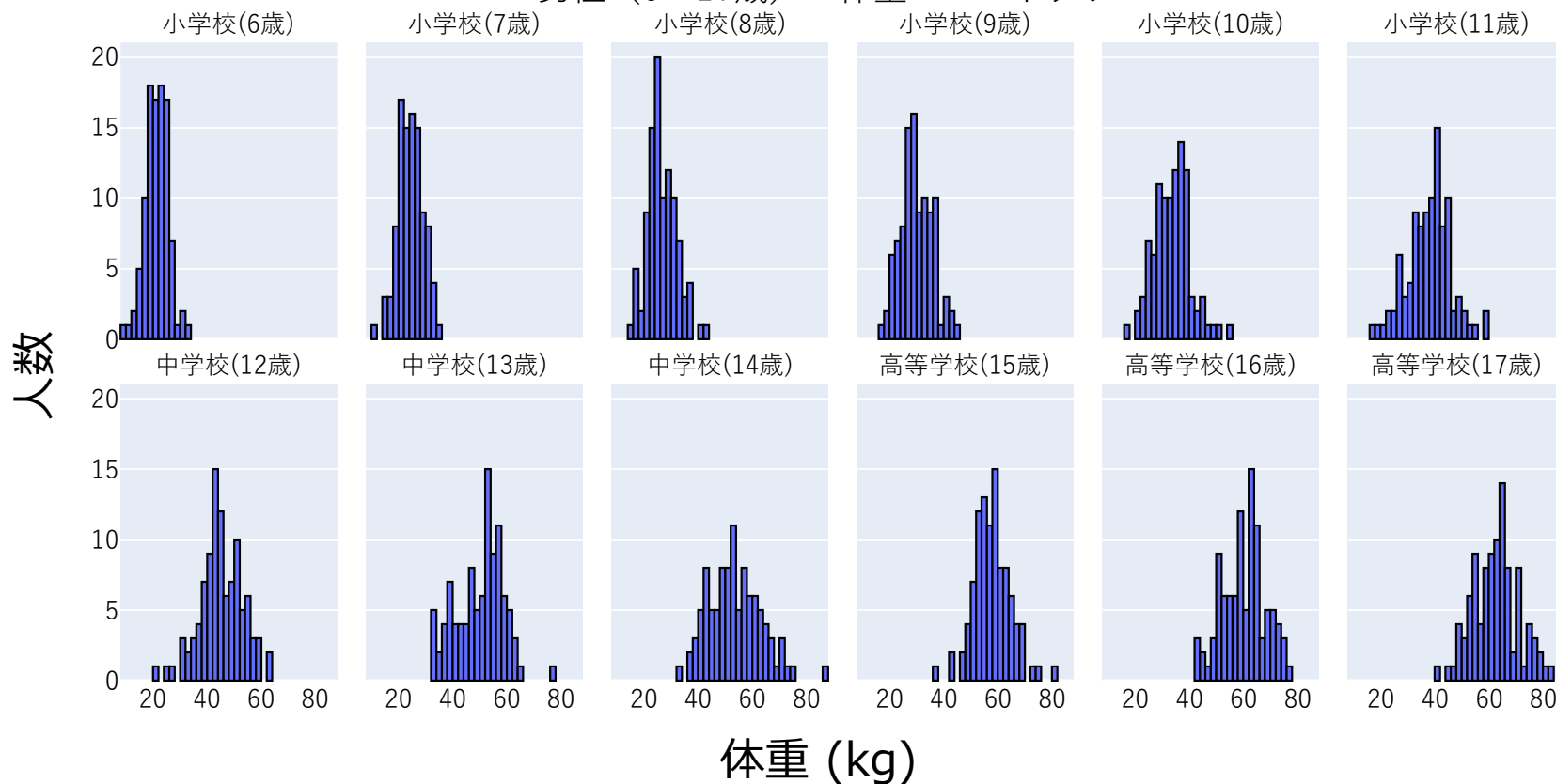


70kgや80kgも多いように見えてしまう

ヒストグラムの比較

- 年齢ごとの体重の「分布」がわかって面白い

男性（6～17歳）の体重のヒストグラム

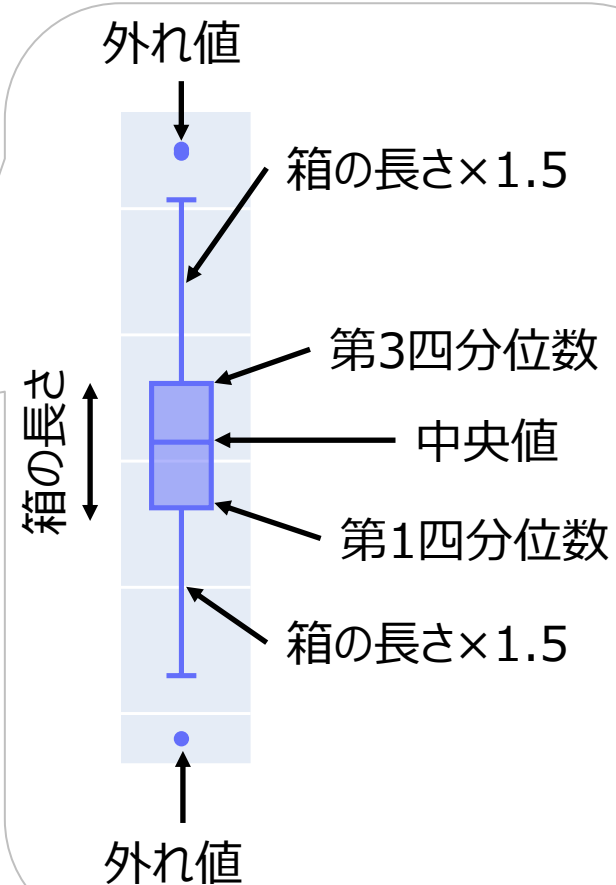
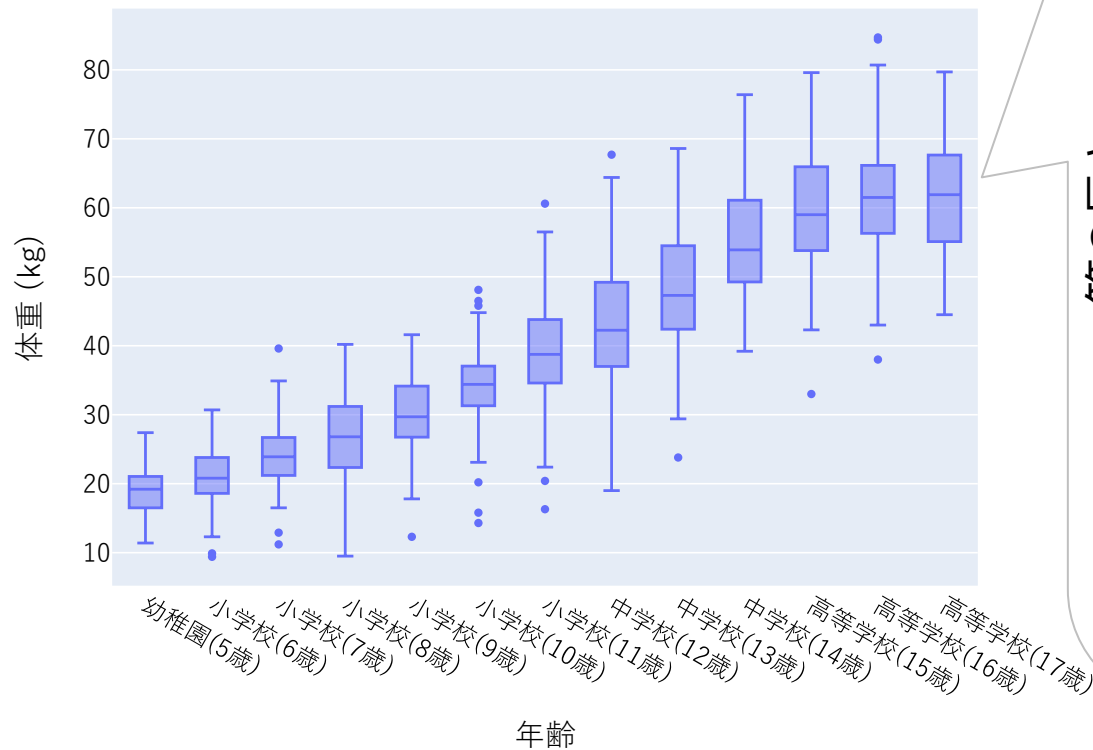


箱ひげ図： 各ビン内でのデータ分布をより詳細に

- 中央値・四分位数で簡易的に表現した図

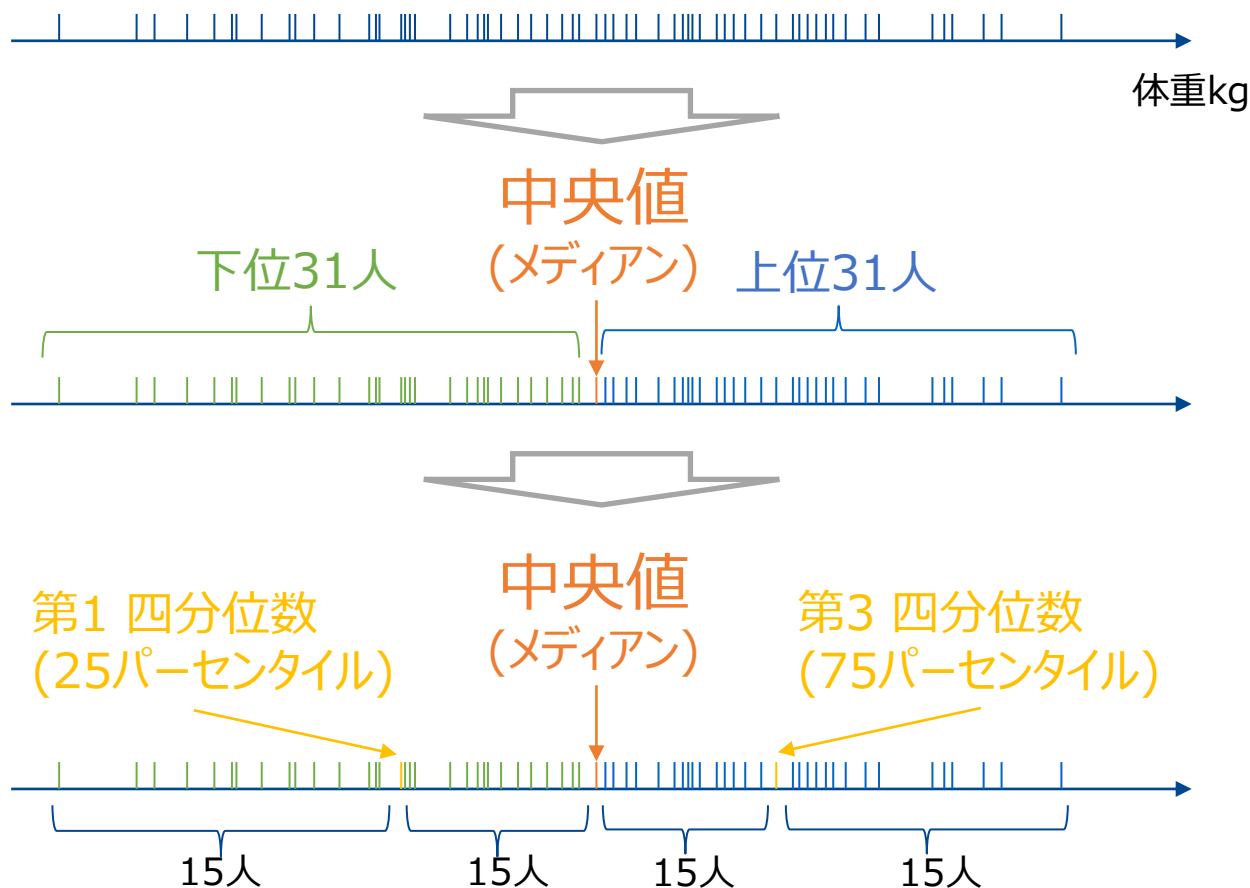
次スライド

男性（5～17歳）の体重の箱ひげ図



中央値・四分位数

あるビンでの体重データ(この例では63人分)



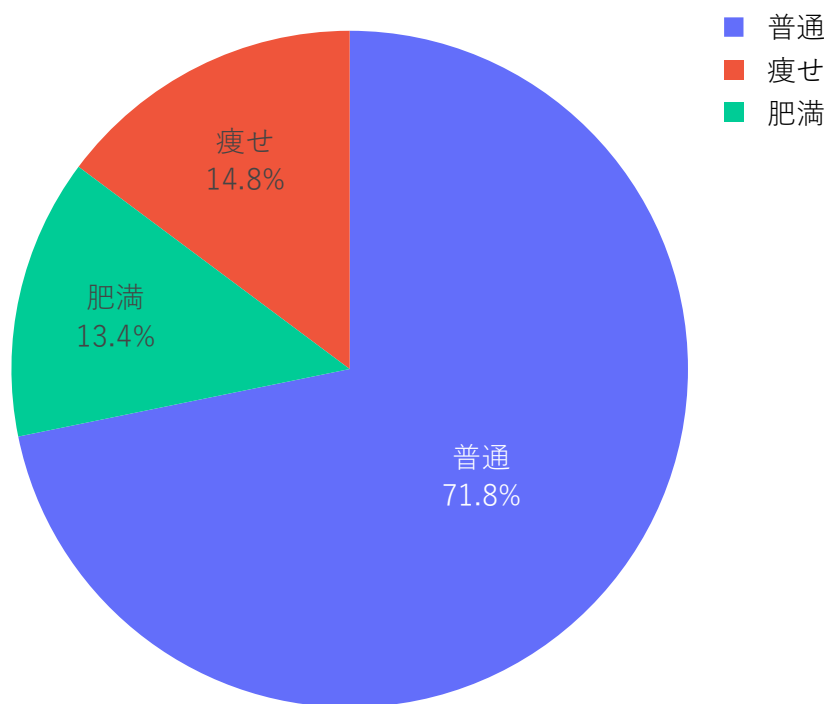
パイチャート

いわゆる「円グラフ」.
科学分野では, あまり使われない

パイチャート

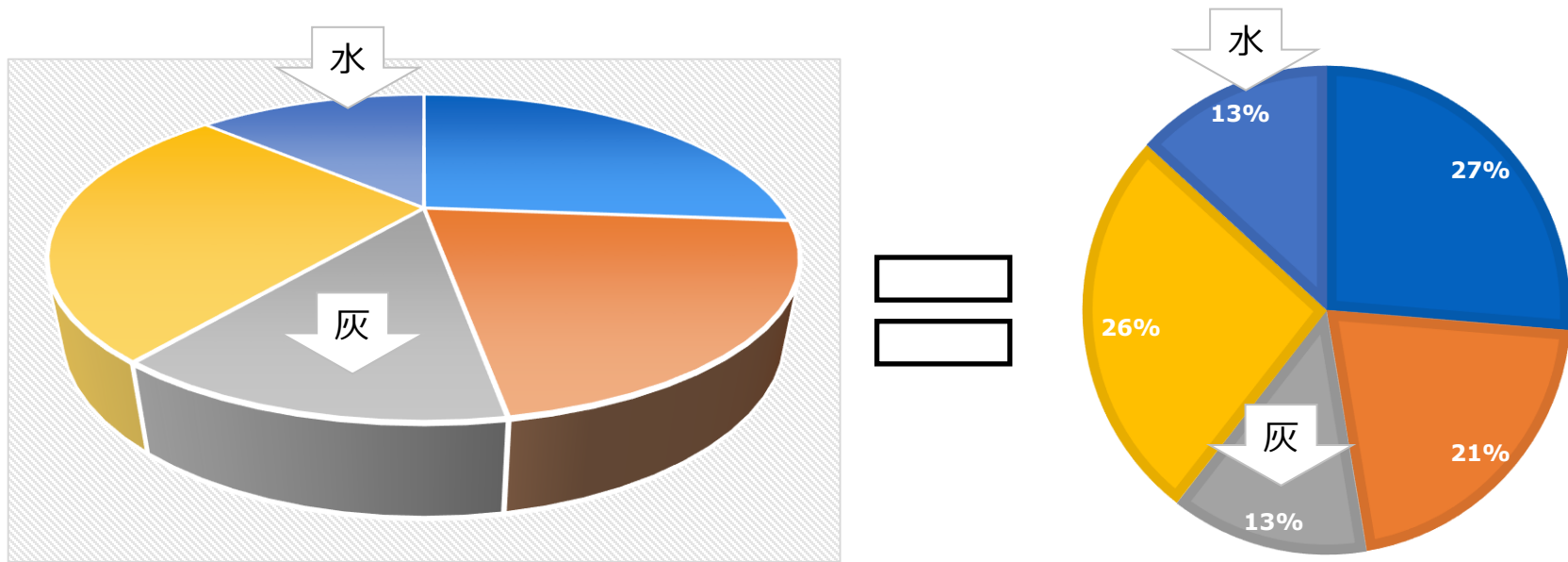
- 円の面積で割合を表示する図
- データ全体に対しての各データの割合を把握する場合に有効
 - 全体的にどのくらい痩せてる人があるのか知りたい！

男子高校生17歳体型 パイチャート



3Dパイチャートの罠： 「わざわざ使う必要はない」と覚えておこう

- 立体的になることで、面積の大小関係が異なるように見える
 - 見栄えはいいが、実際の値がなければ、誤解してしまう



灰色 > 水色 に見える！！

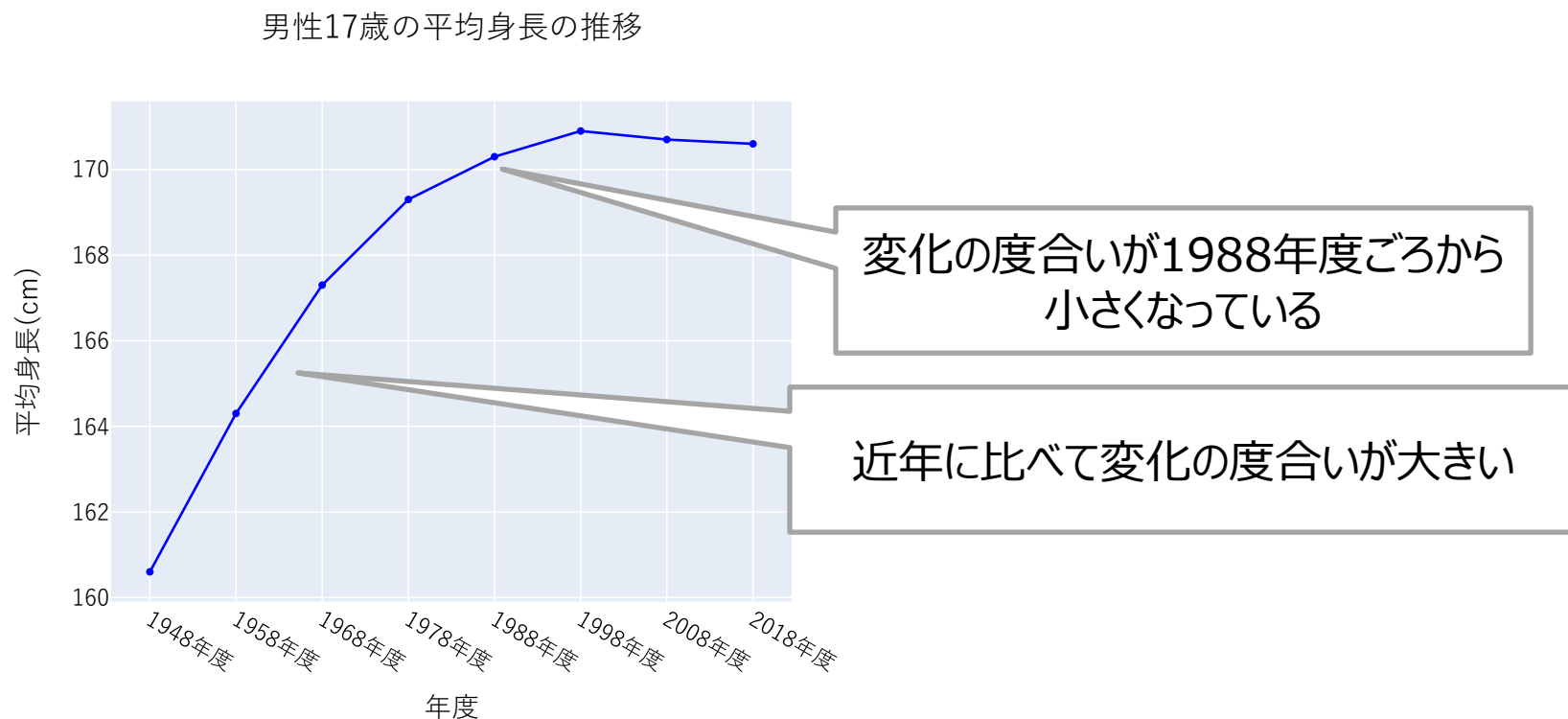
ホントは 灰色 = 水色 ！！

- 実は2Dパイチャートでも微妙な角度差はわかりにくいので、科学や工学の分野では棒グラフを使うことのほうが多い

折れ線グラフ

折れ線グラフ

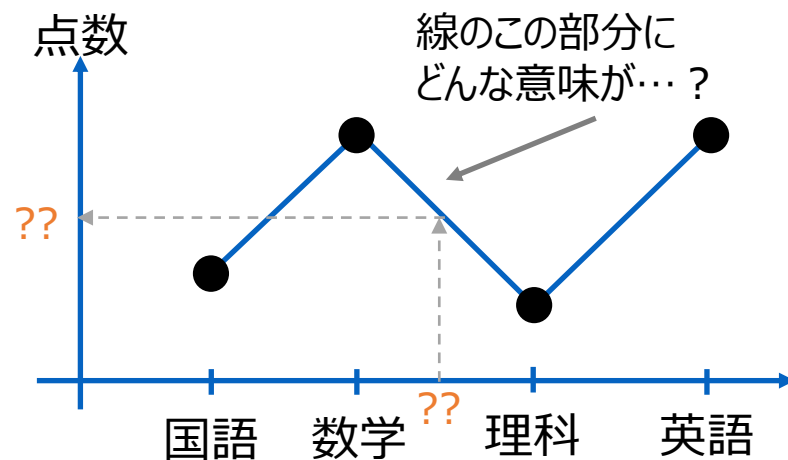
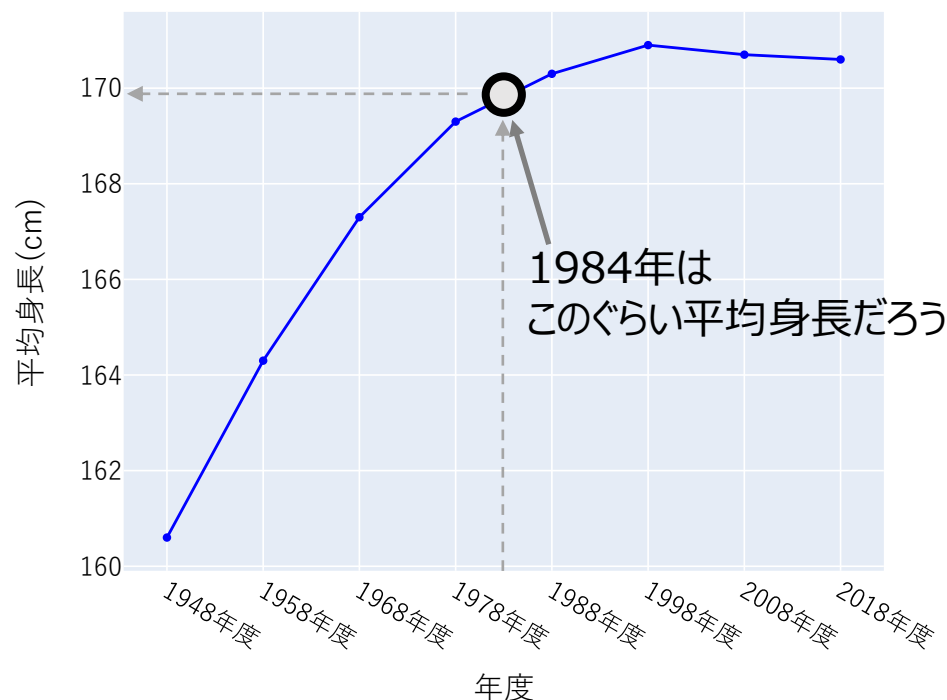
- データ点の間を直線で結んだ図
- データ点の間の変化の傾向を把握する場合に有効
 - 平均身長の推移を知りたい



折れ線グラフの線の意味

- 線は観測されなかった点と点の間を補間する意味合い
- 関連のないデータを線で結ぶのは不適當

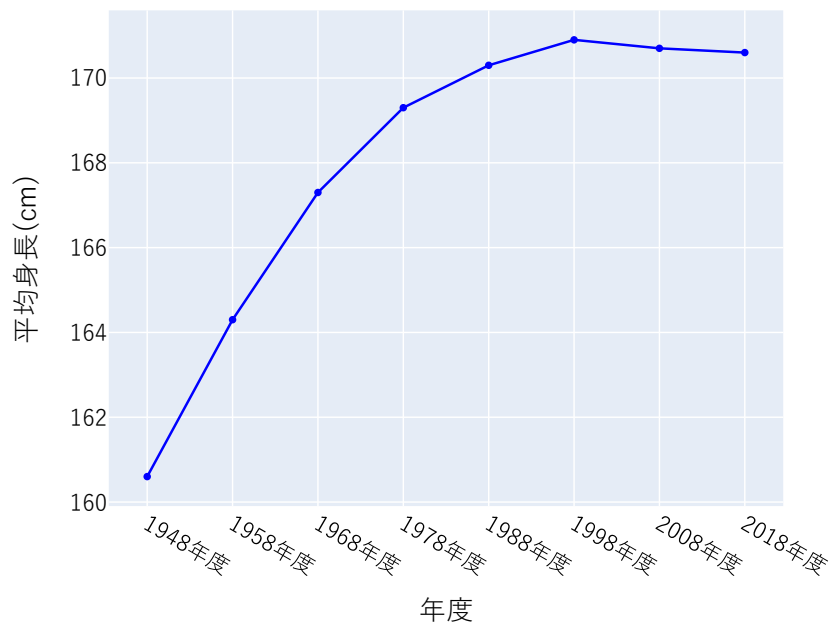
男性17歳の平均身長推移



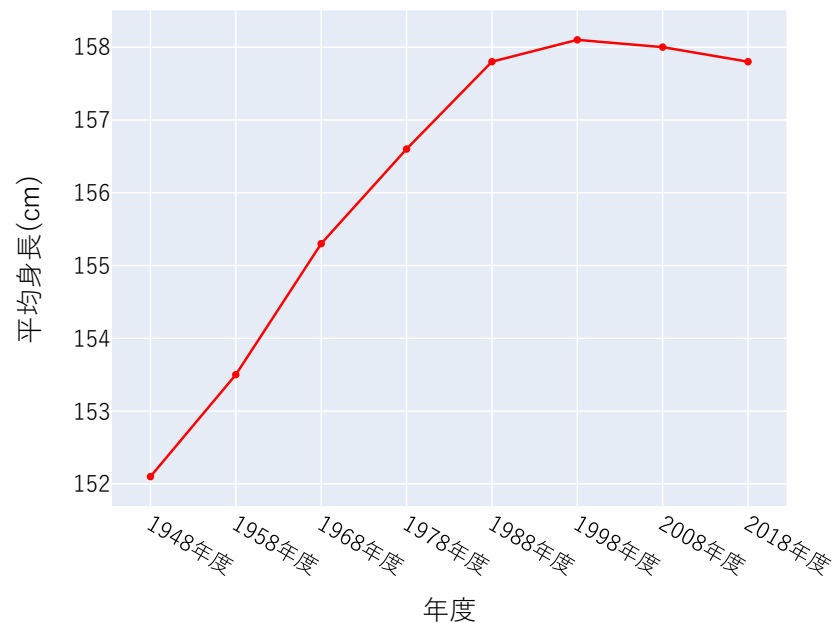
折れ線グラフの注意

- 折れ線グラフを比較する際には**縦軸**の範囲や幅を同じにする

男性17歳の平均身長推移



女性17歳の平均身長推移

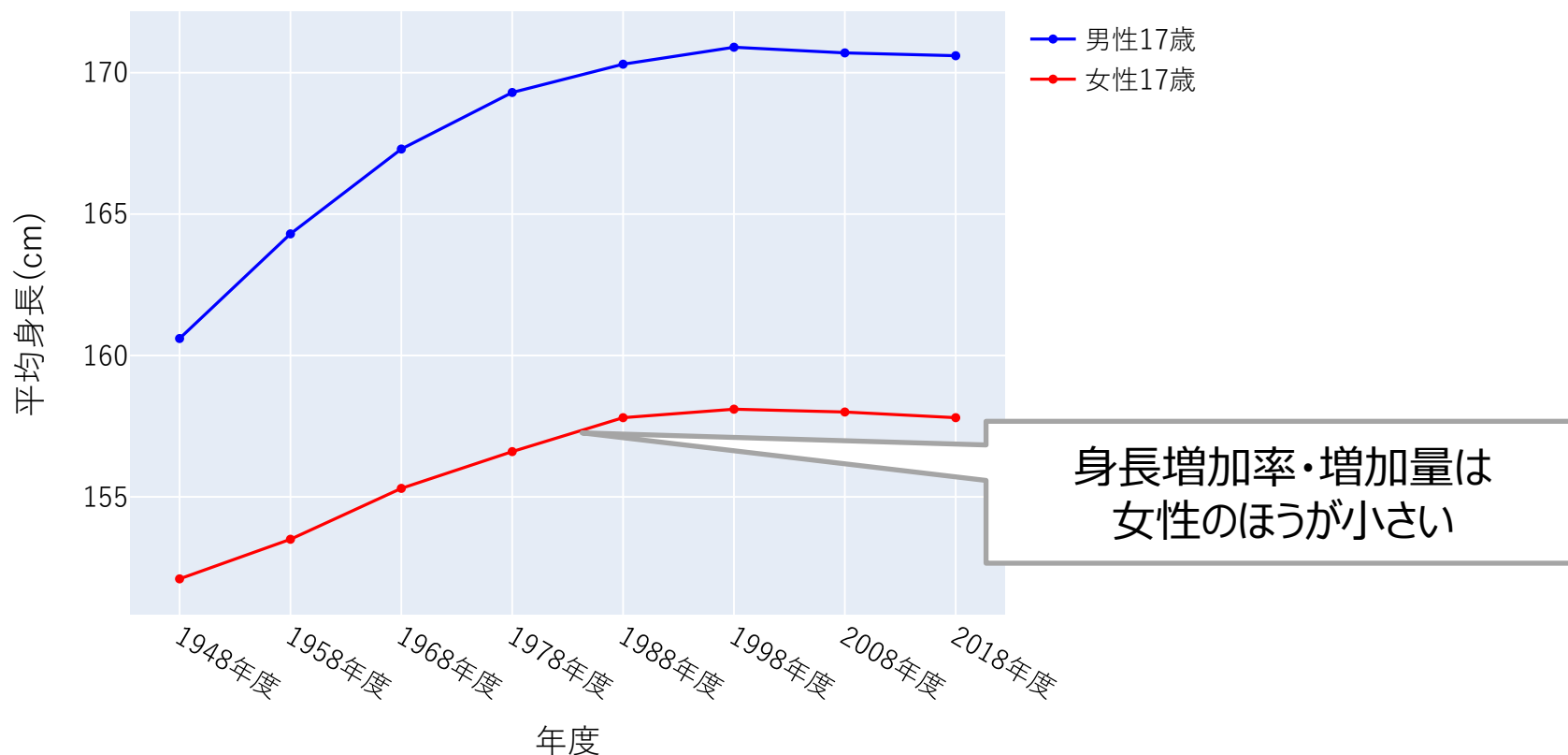


一見すると、男性女性で同じ変化に見えるが、**縦軸の範囲や幅が異なる！！**

折れ線グラフの比較

- 同じグラフ描画内に描画するとよい

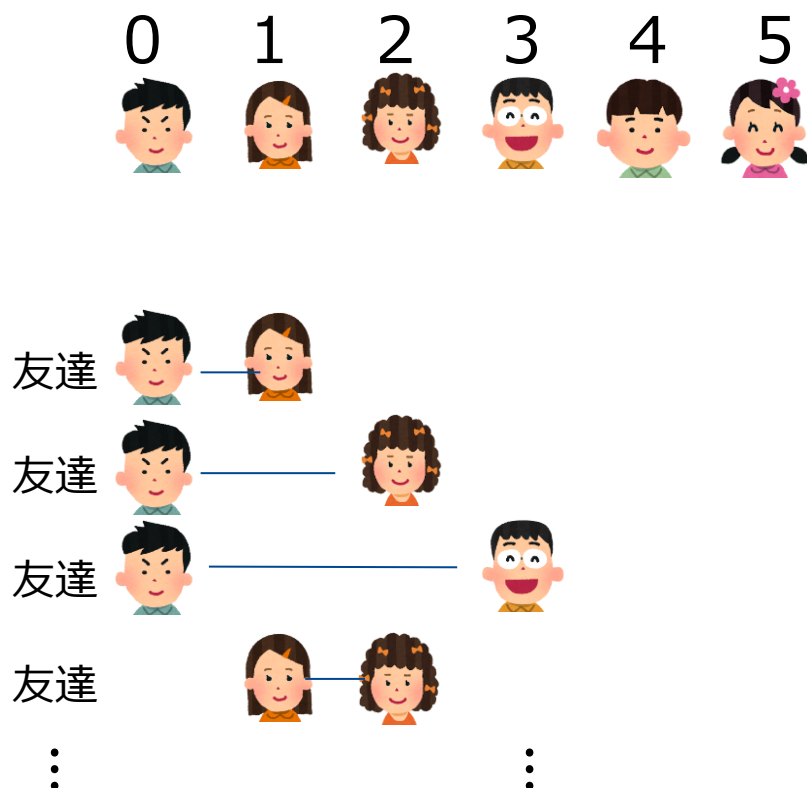
平均身長推移



グラフ（ネットワーク）による 「関係」の可視化

お友達関係の可視化

- SNSのお友達関係を可視化したい



友達関係を表で表す

- 「友達なら1」「友達でないなら0」を意味する表で表現
- 例：Zachary's karate club 34人の空手クラブの交友関係

	0	1	2	3	4
0	0	1	1	1	1
1	1	0	1	1	0
2	1	1	0	1	0
3	1	1	1	0	0

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33					
0	0	1	1	1	1	1	1	1	1	0	0	0	1	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0					
1	1	0	1	1	1	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0	1	0	1	0	0	0	0	0	0	0	0	1	0	0					
2	1	1	0	1	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1				
3	1	1	1	0	1	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
4	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
5	1	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
6	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
7	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
8	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1				
9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1				
10	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
12	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0				
13	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1				
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1			
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1			
16	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
17	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0			
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1		
19	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1			
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1		
21	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1		
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1		
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	
27	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1		
28	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
29	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
30	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
31	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	
32	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1
33	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	1	1	0	0	1	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	

3さんと4さんは友達ではない

さんと0さんは友達

3さんと4さんは
友達ではない

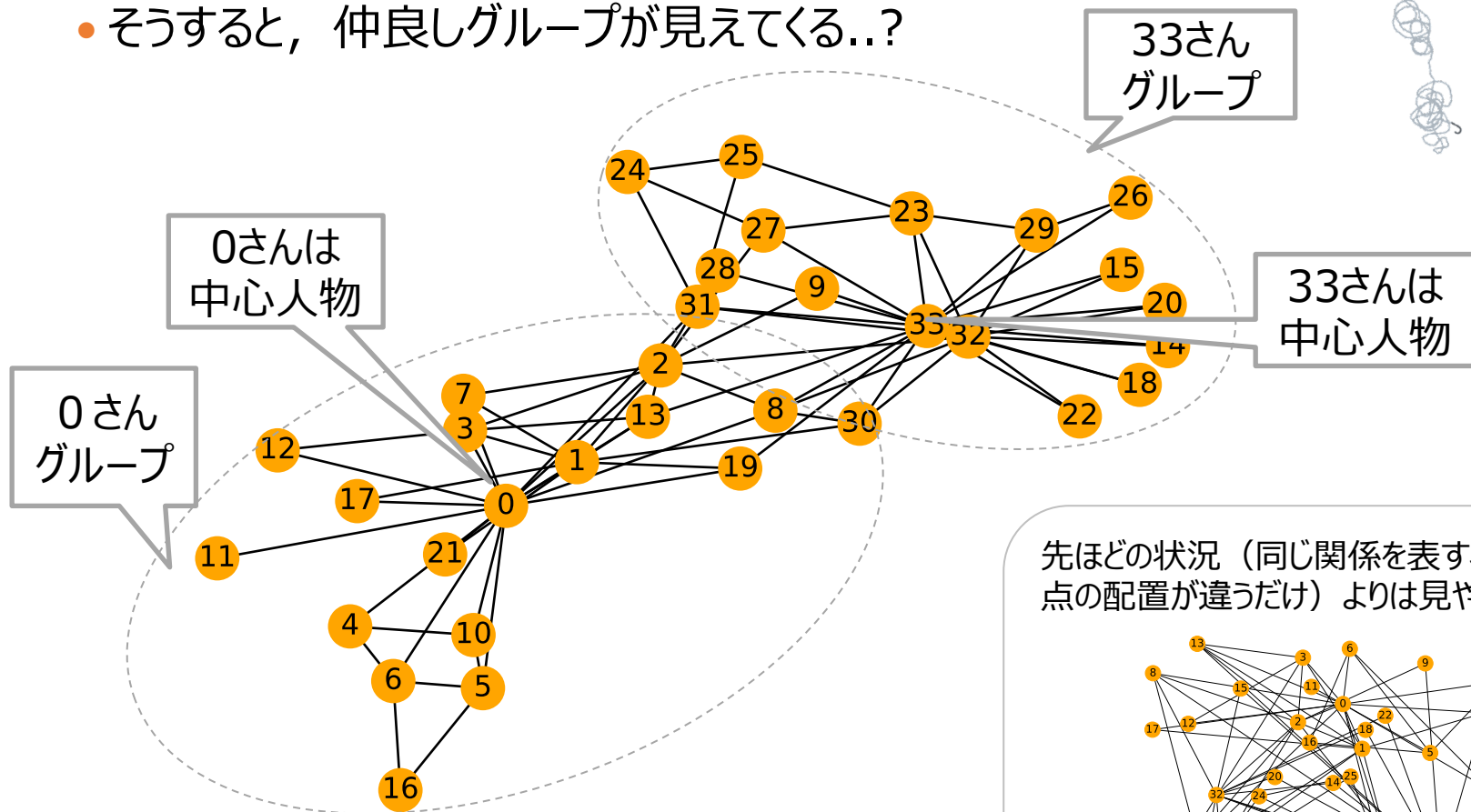
3さんと0さんは友達

- 折れ線グラフや
棒グラフの「グラフ」
とは意味が違うので
注意

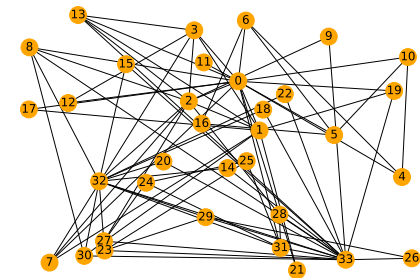


ネットワークの点（ノード）はどこに置いてもOK. そうなら、点の配置の工夫を見やすくしたい

- 例：友達は近くに配置，それ以外は遠くに配置する
 - そうすると，仲良しグループが見えてくる..?

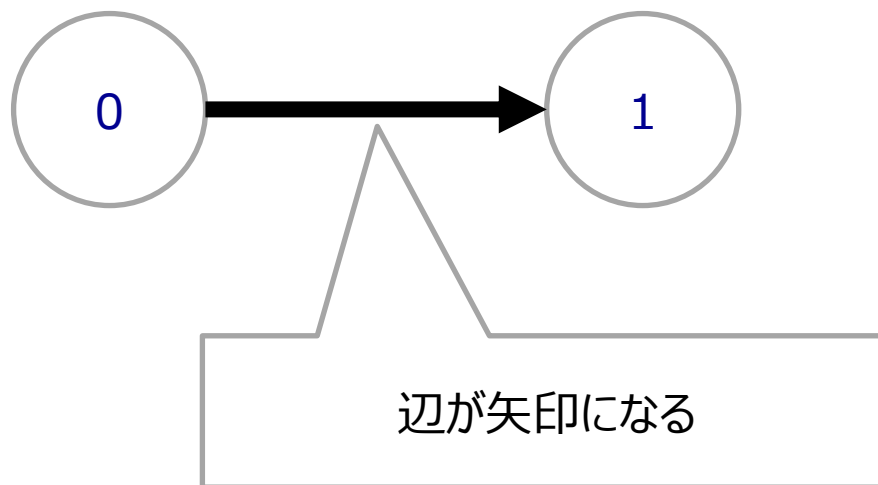


先ほどの状況（同じ関係を表すネットワーク、点の配置が違っただけ）よりは見やすい？



有向グラフ

- 辺に向きがあるネットワーク
- 0さんは1さんを友達とってるけど、1さんは思っていない場合



有向グラフ+色+地図

- ノードで場所, エッジで人の移動**方向**を表現
- 色で人流の量を表現
 - 人が少ない【青 < 緑 < 赤】 人が多い



食堂からの
行き来が多い

人の流れの可視化

散布図と二次元ヒストグラム

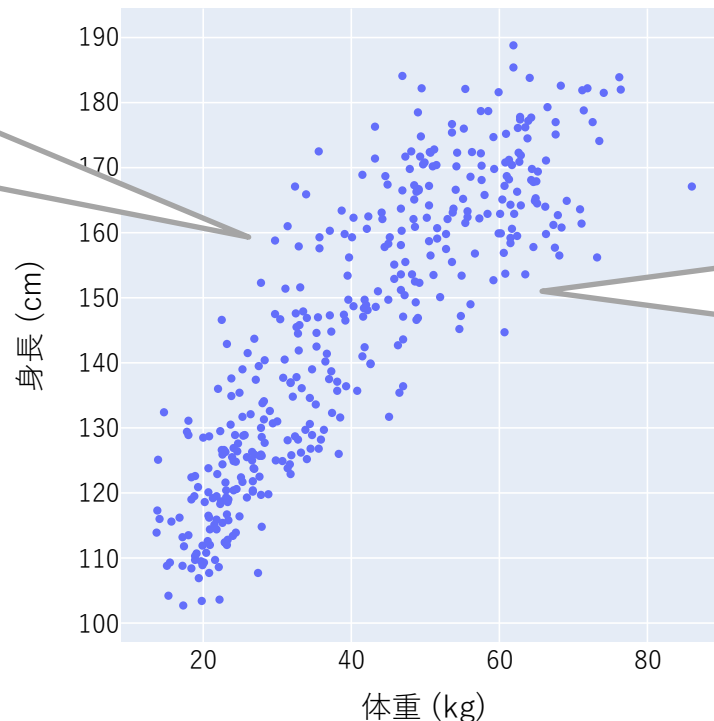
分布や広がり of 可視化

散布図

- 平面上に点の集合としてデータを表現した図
- 2種類のデータの傾向を同時に把握したい場合に有効
 - 体重と身長傾向を知りたい

男性（5～17歳）の体重と身長散布図

青色1点が
(体重, 身長) の
座標を表している

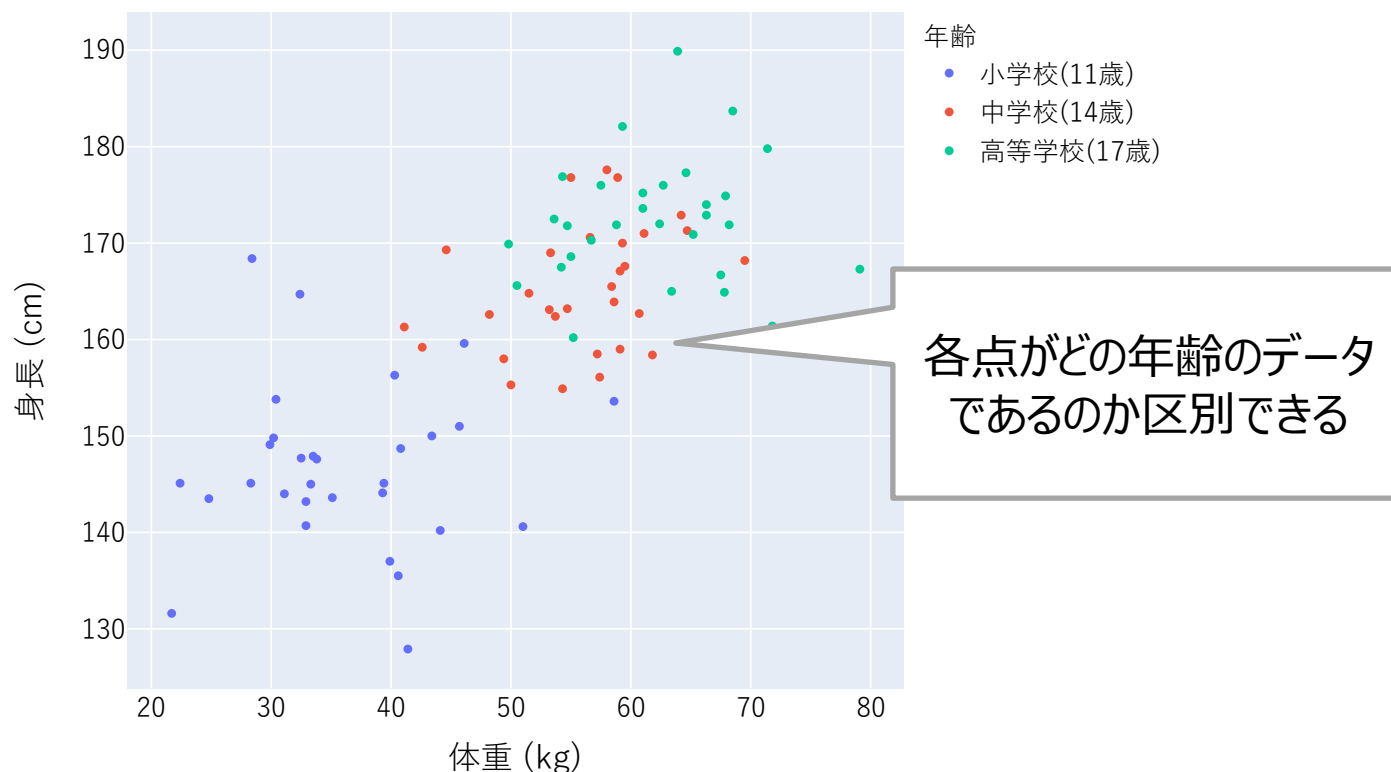


右肩上がり関係が
ありそうだ

色 + 散布図

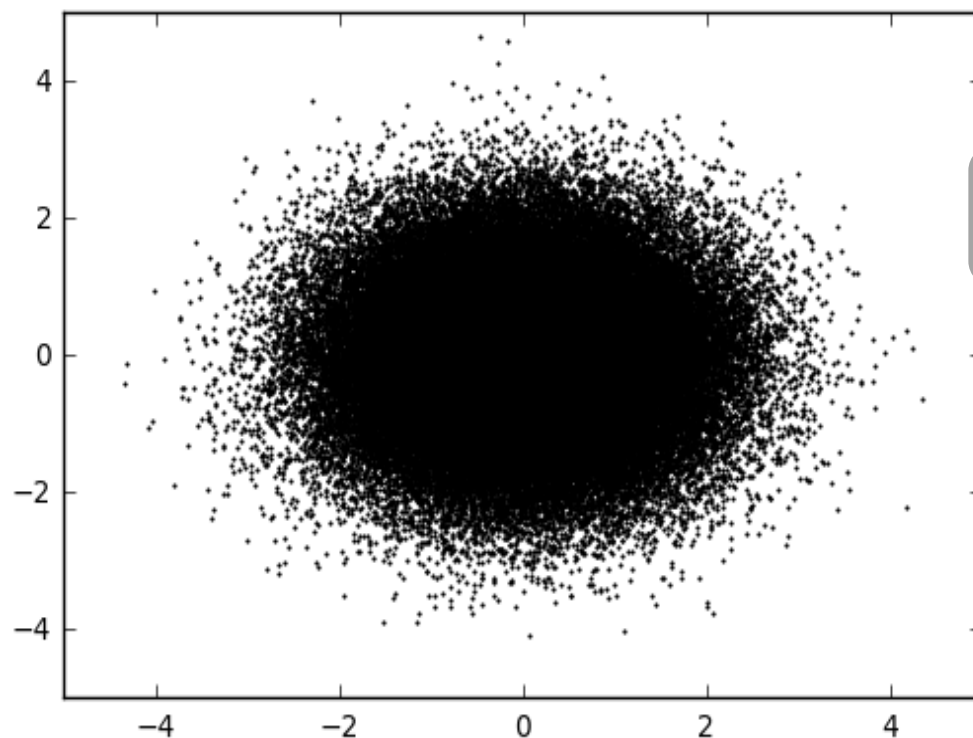
- 色によって3種類の値を可視化可能
 - 形や点の大きさを変えることでさらに多くの情報を同時に可視化できる

男性（最終学年）の体重と身長散布図



データ数が膨大になると、散布図は不向き

- 点が重なってよくわからない…

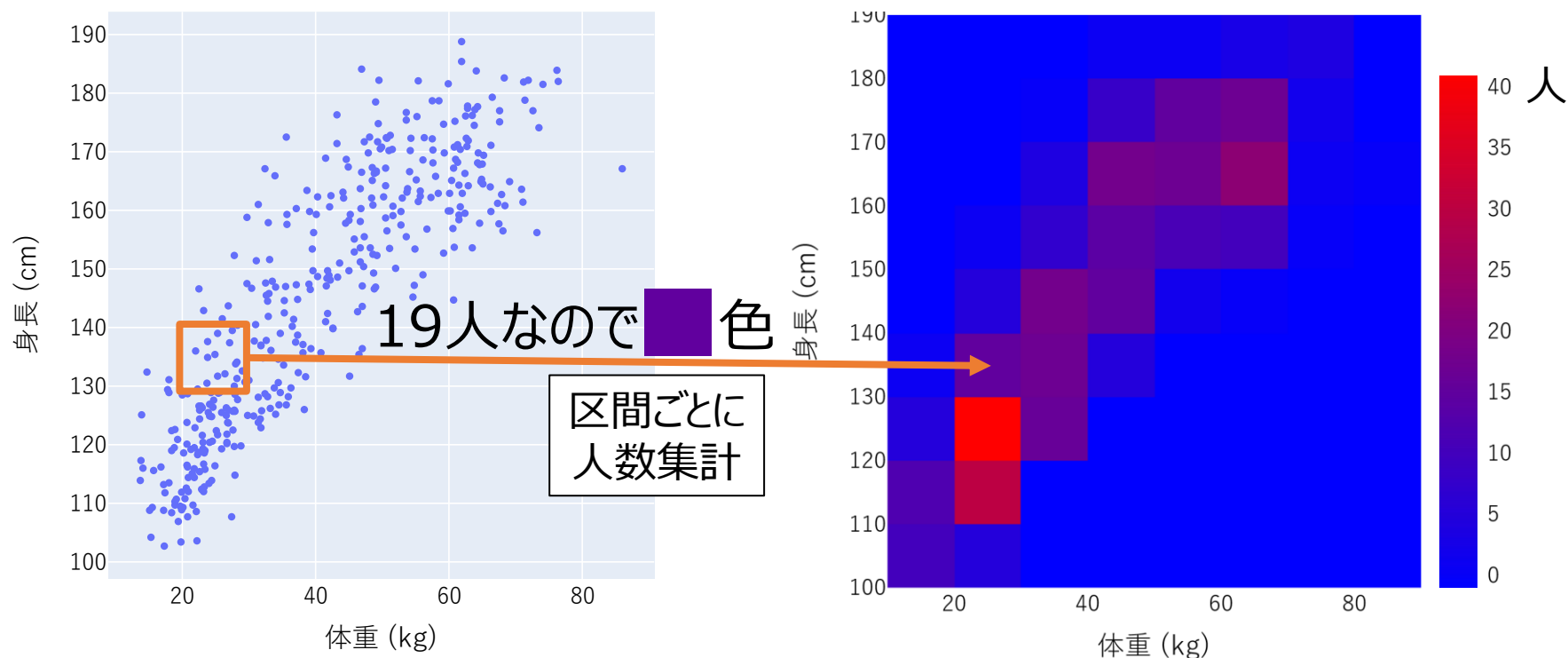


真ん中に行くほど高密度なのか、
それとも中心付近は一定密度なのか…



ヒートマップ（2次元ヒストグラム）： 色によって頻度(度数)を表現した図

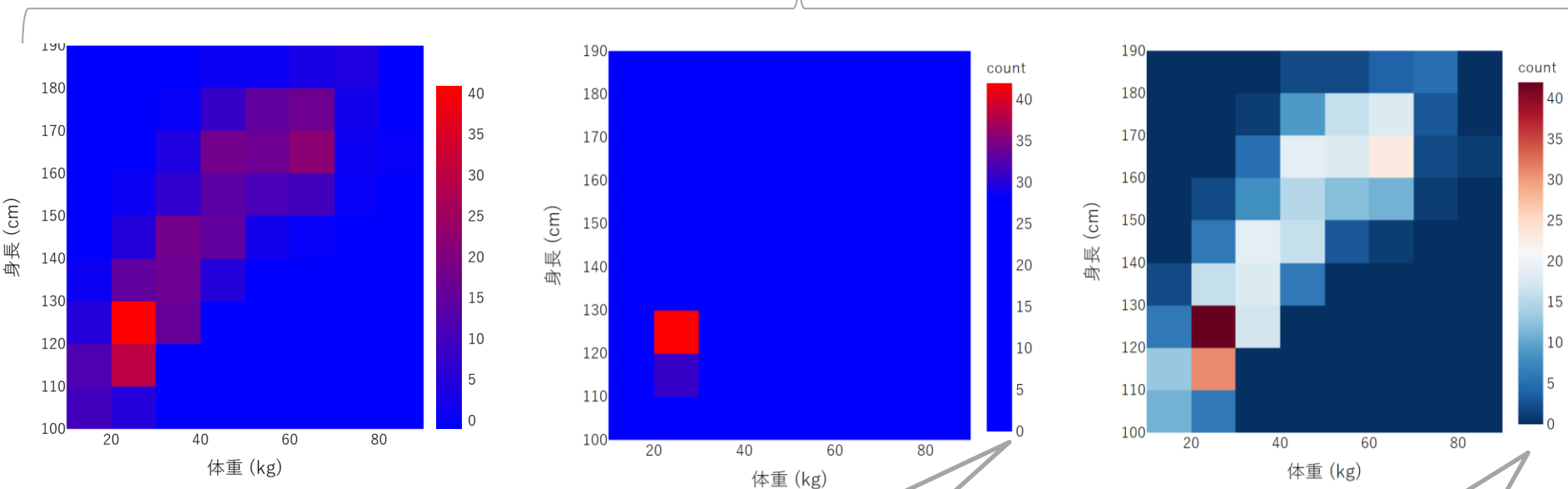
- 区間に分け、その中の点数に応じた色で、その区間を表示
- データ数が多くなっても大丈夫！ ざっくりと傾向をつかむことが可能に
 - ただし、細かい状況はわからなくなる。さらに…(次スライド)



ヒートマップの色には気を付けよう！ 色の設定で同じデータでも印象・見方が大きく異なる

- なので、色設定によっていくらでもズル(?)できてしまう

全部同じデータだが...

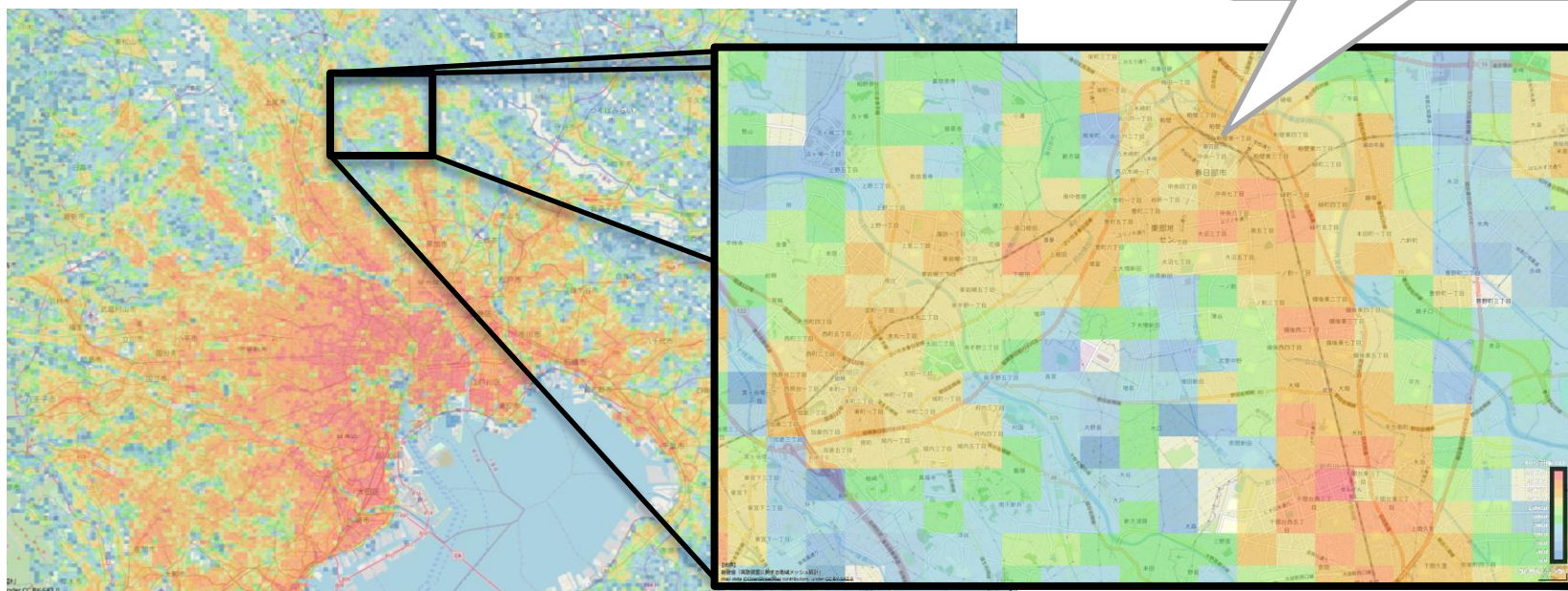


最大値付近のみが赤色
その他の区間は0に見えてしまう

中央値を境に大きい値は赤系
小さい値は青系

ヒートマップ+地図： 人口密度をヒートマップで可視化

東京近郊の人口密度



拡大

「RESAS（地域経済分析システム）-人口メッシュ-」

<https://resas.go.jp/population-mesh/#/map/13/13101/2/2015/0/0/0/13.699533624450218/35.95712520463781/139.75173925531294/2015/0/0>

<https://resas.go.jp/population-mesh/#/map/13/13101/2/2015/0/0/0/10.23602789302161/35.8295917737677/139.93371656011604/2015/0/0>

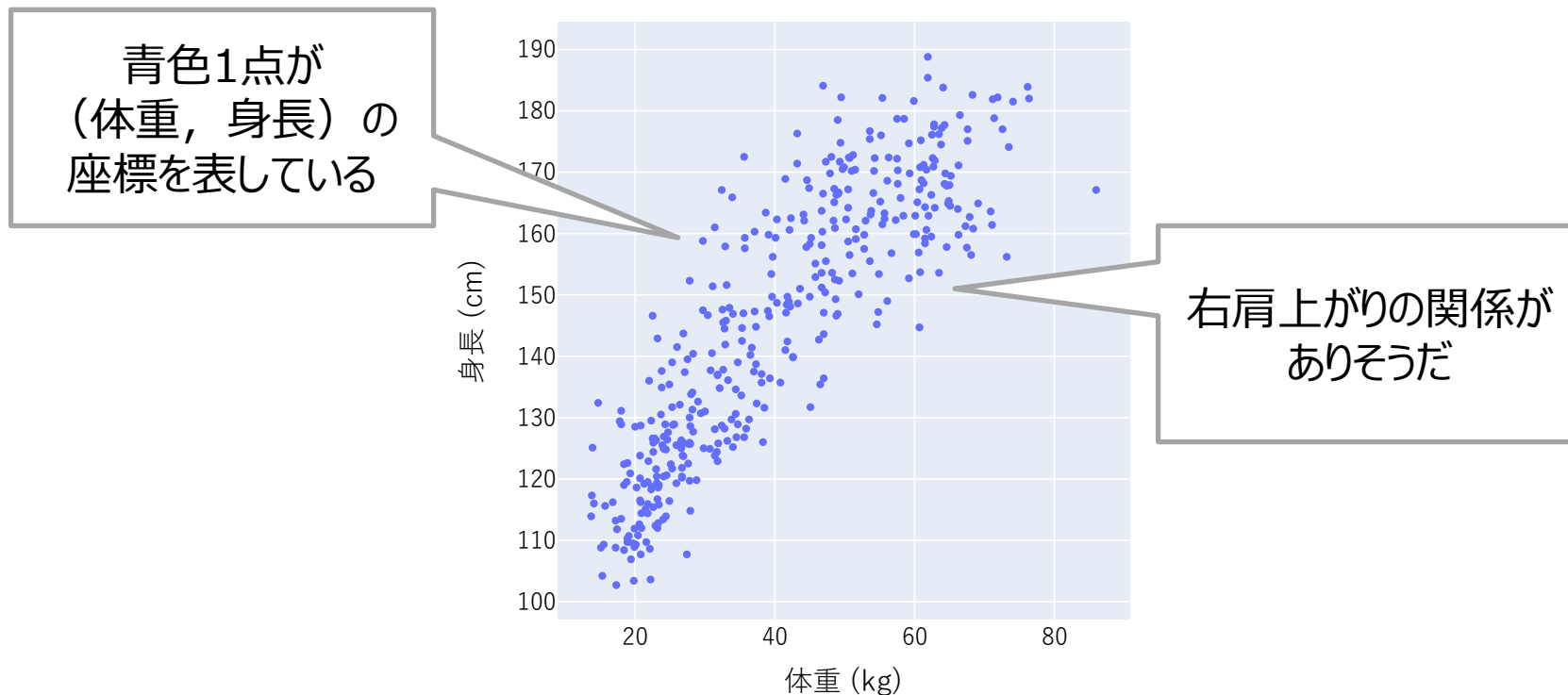
を加工して作成

多次元データの可視化

(再び) 散布図

- 以下の例のように、1 データ = 2 次元（体重，身長）なら，散布図がベスト

男性（5～17歳）の体重と身長散布図



- でも、もし 1 データ = 4 次元（身長，体重，腹囲，足の長さ）だったら？！？

3種類以上の要素の関係の可視化

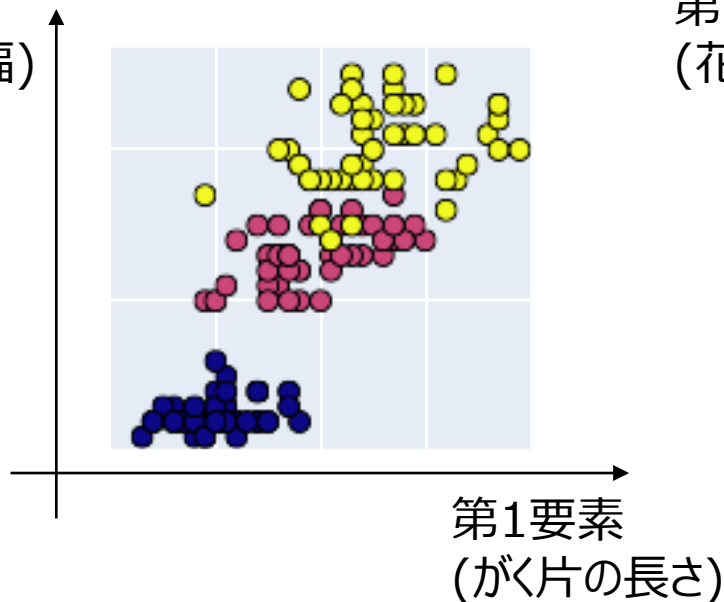
- 3種類以上の要素は直接表示できない
- 3種類以上の要素を持つデータをまとめた散布図でもデータ分布を表現できていない

「3つだったらできる」と思うかもしれませんが、よく考えると、正確には無理なことがわかるはず

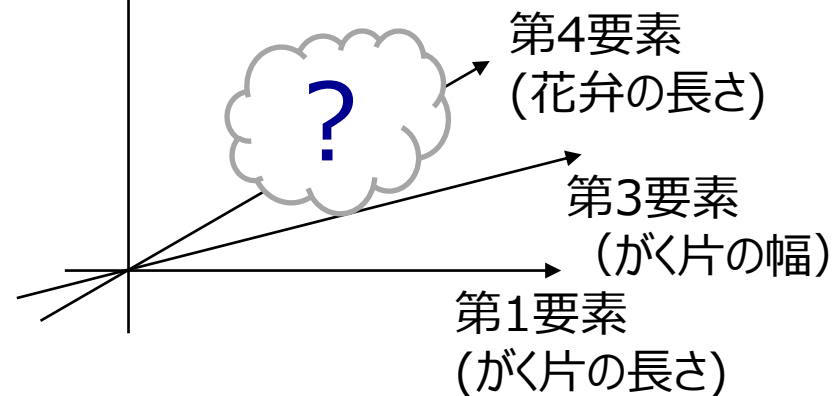
2種類の要素は描ける

3種類の要素は平面に正確には描けない

第2要素
(花弁の幅)

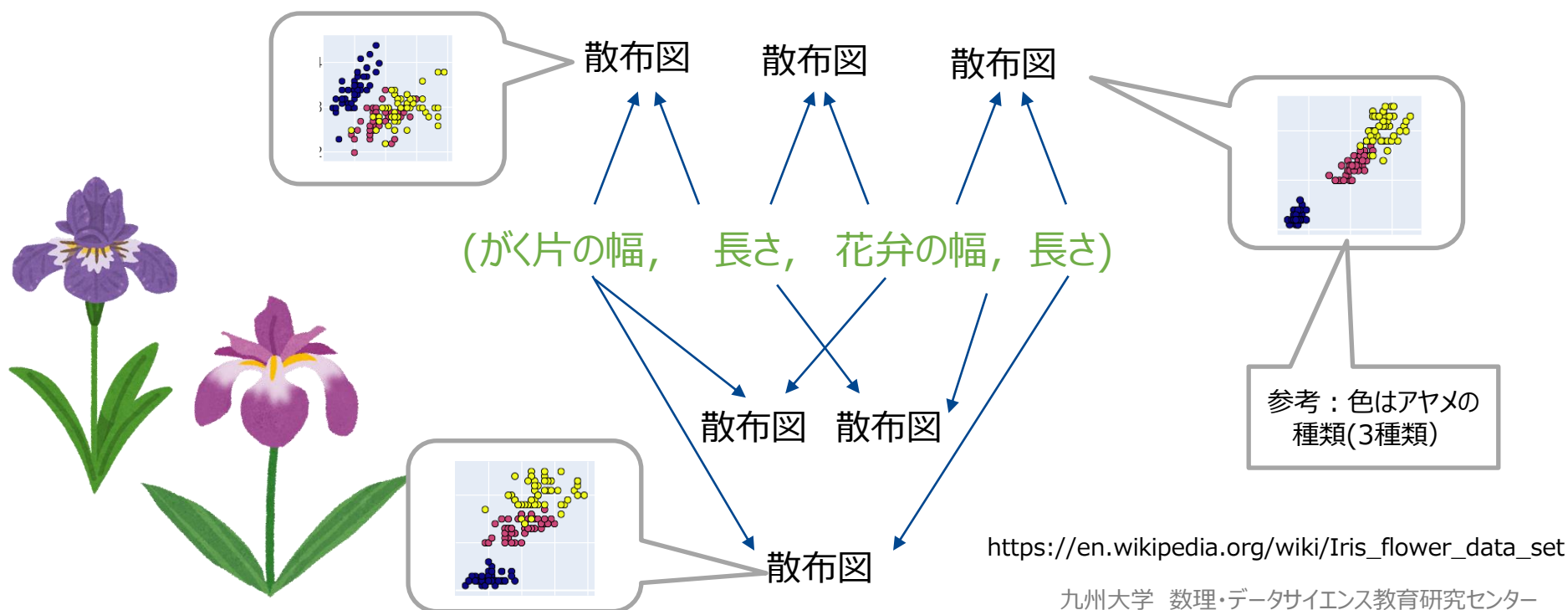


第2要素
(花弁の幅)

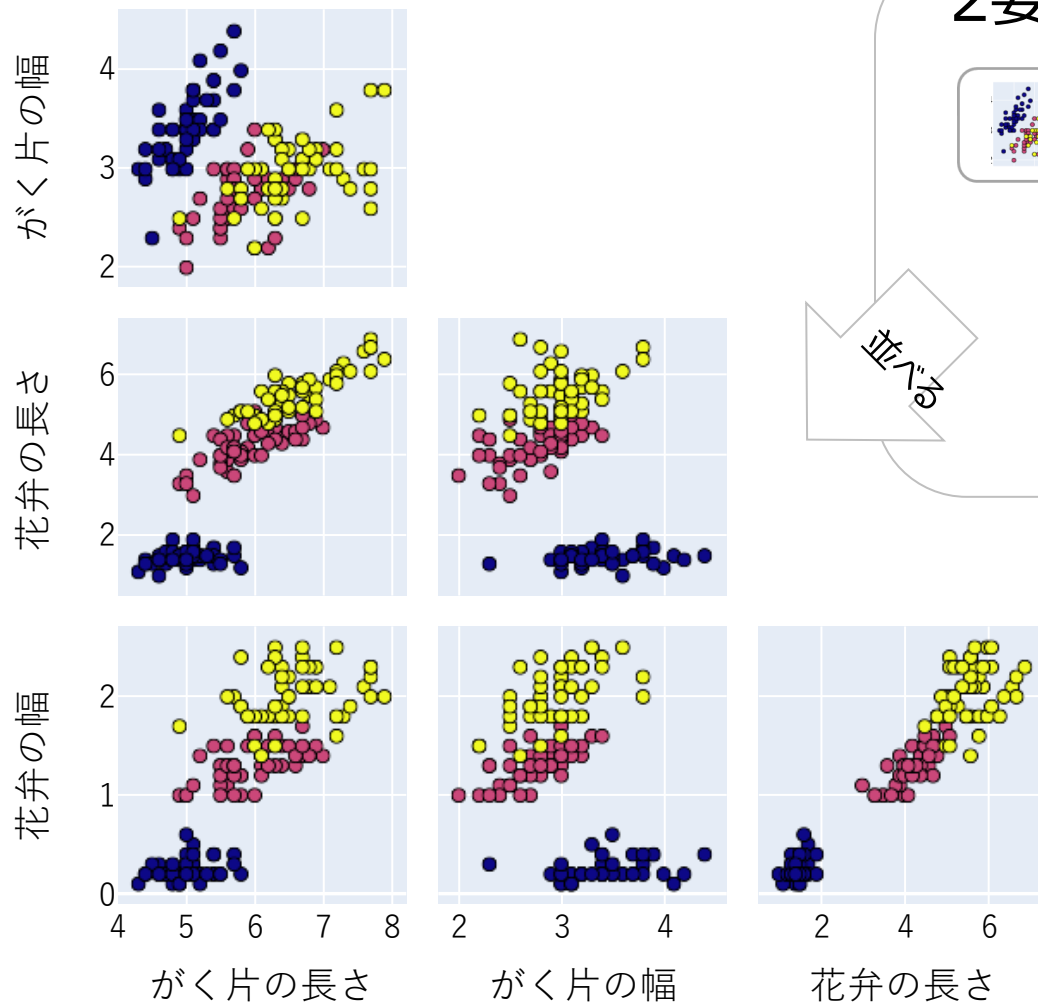


要素数を削って可視化 (1/2)

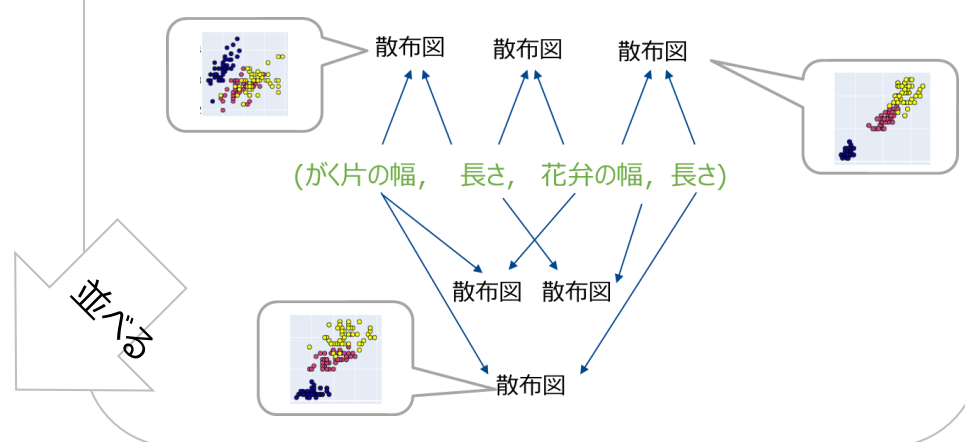
- 2種類の要素なら散布図で可視化
- ならば、要素をふたつずつピックアップしてそれぞれ散布図を描くと？
 - アヤメの花の形（4次元データ）の場合 → 6つの散布図



散布図行列： 複数の散布図を並べたもの



2要素ずつ→合計6散布図



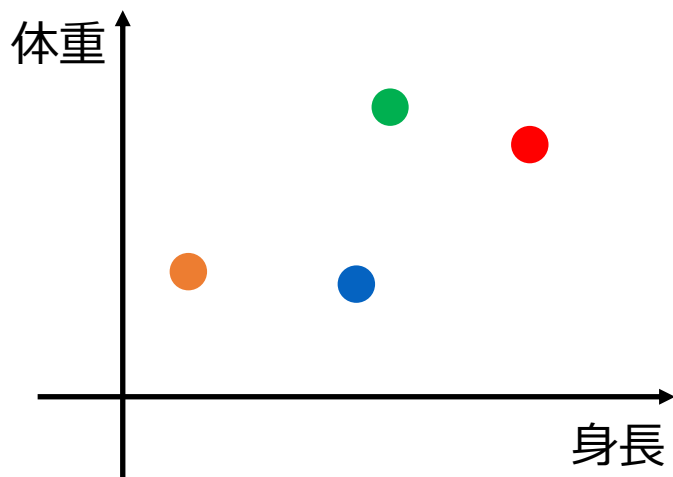
要素数を削って可視化 (2/2)

- 多要素のデータを2要素のデータに変換(?)

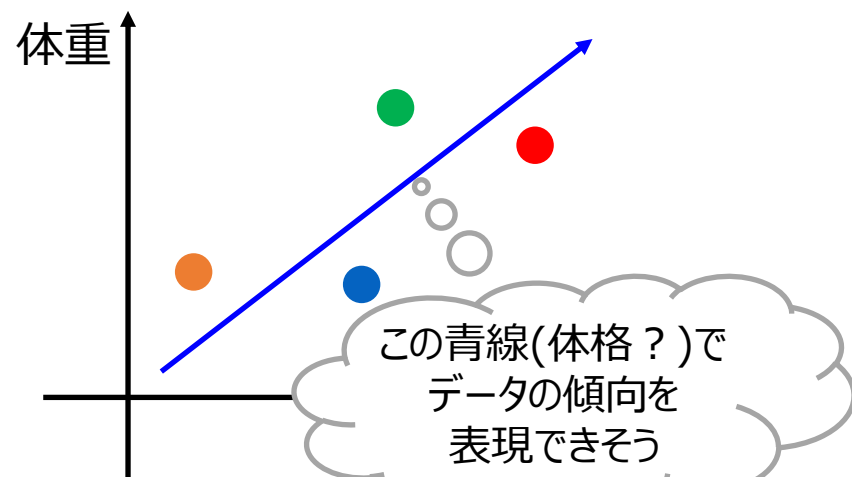


変換って何？ そんな魔法のようなことができるのか？

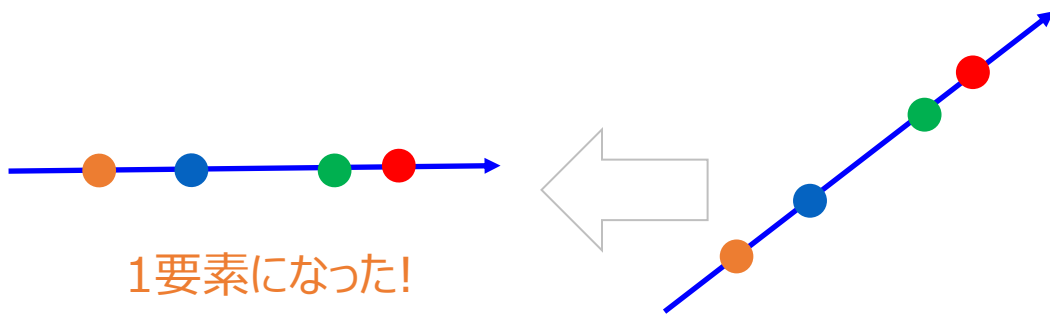
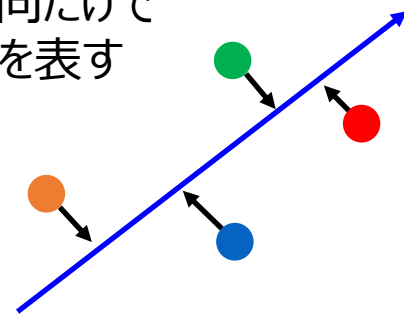
データの変換：2要素→1要素の場合



各データは(身長, 体重)の2要素



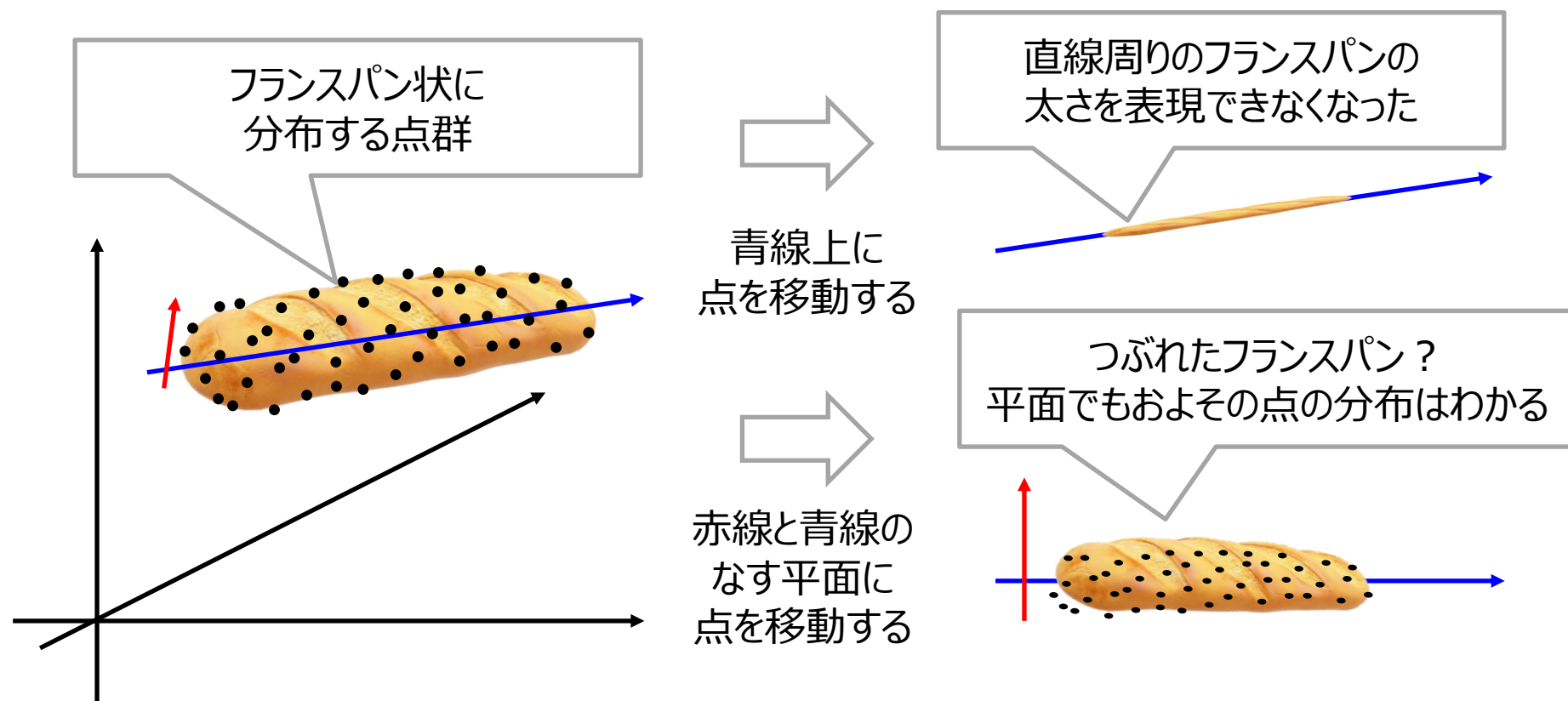
青線方向だけで
データを表す



1要素になった!

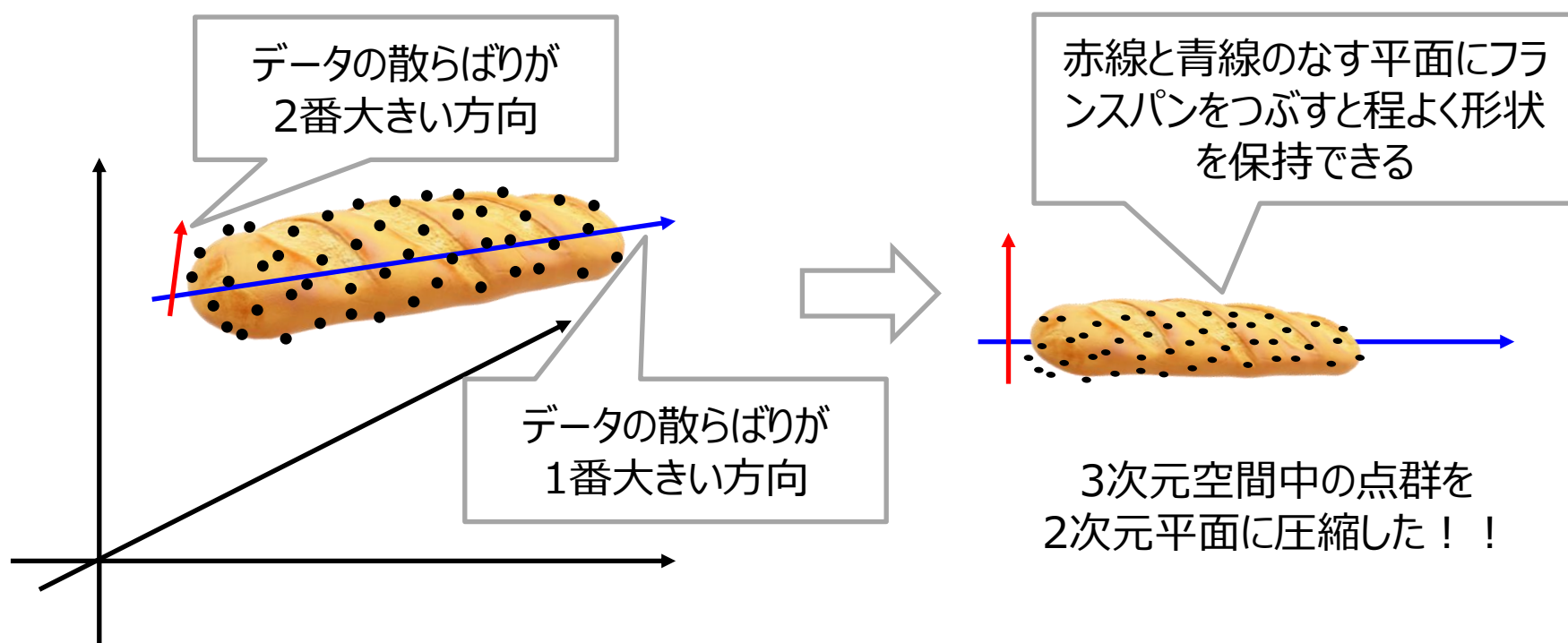
データの変換：3要素→1要素の場合

- フランスパン状に分布する点群を直線・平面で表現してみる
 - 別にフランスパンでなくてもよいのですが…



データの散らばりが大きい方向に注目して、要素を減らす

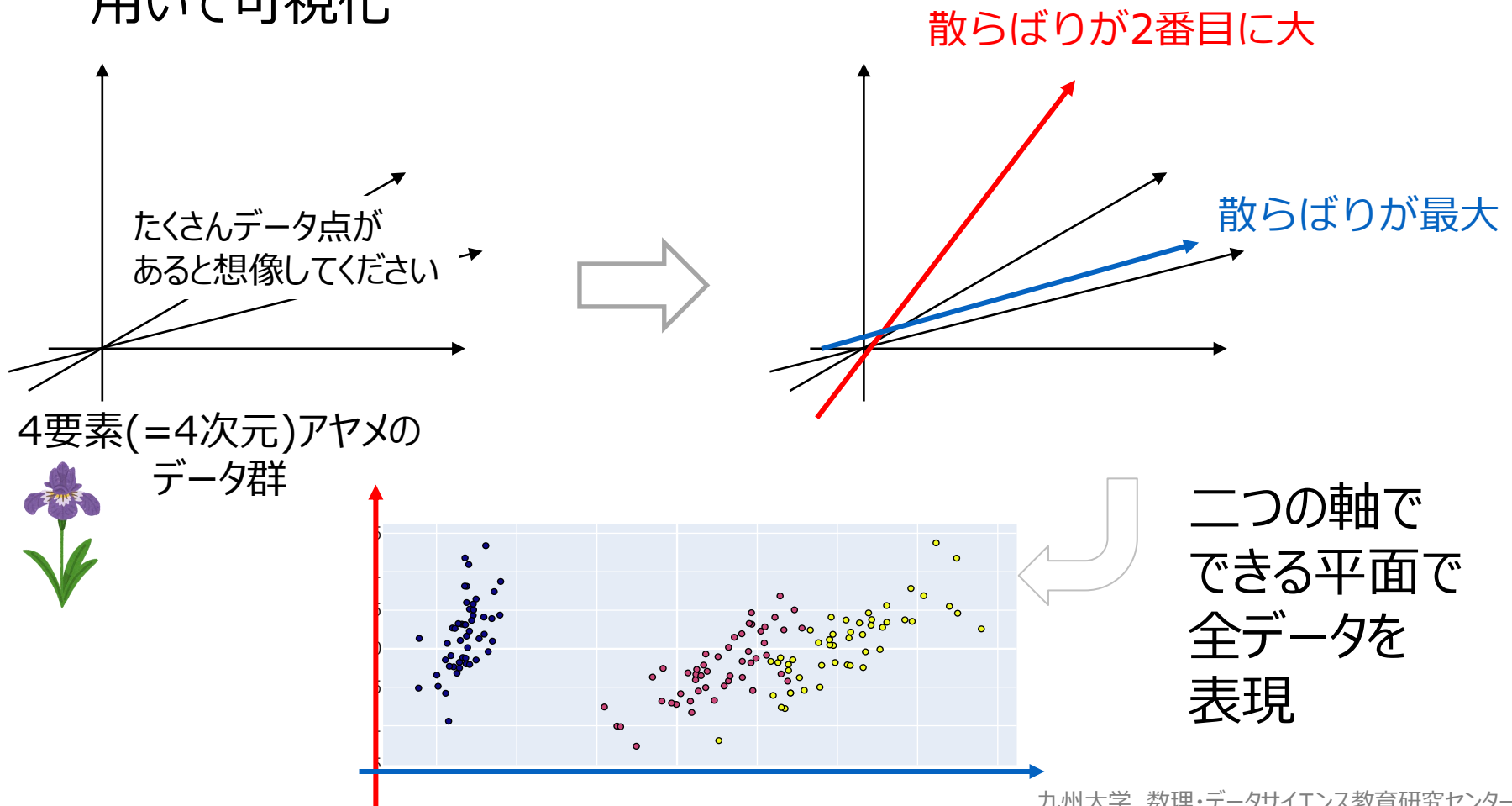
- 散らばりが大きい＝データを表現するために大事



このような考え方を「主成分分析」と呼ぶ

というわけで、要素数を2次元に変換することで、 散布図として可視化することが可能

- データの散らばりが大きい2つの方向を見つけて、それらを平面として用いて可視化



可視化のまとめ

可視化手法まとめ

- データの分布を調べたい→ヒストグラム
- データ分布の比較したい→箱ひげ図
- 数値データの比較したい→棒グラフ
- データ全体に対しての各データの割合を把握したい→パイチャート
- データ点の間の変化の傾向を把握したい→折れ線グラフ
- 2種類の要素の傾向を同時に把握したい→散布図
- 散布図ではデータ点多すぎる→ヒートマップ
- データ間の関係性・接続を把握したい→ネットワーク、有向グラフ



まとめ

可視化

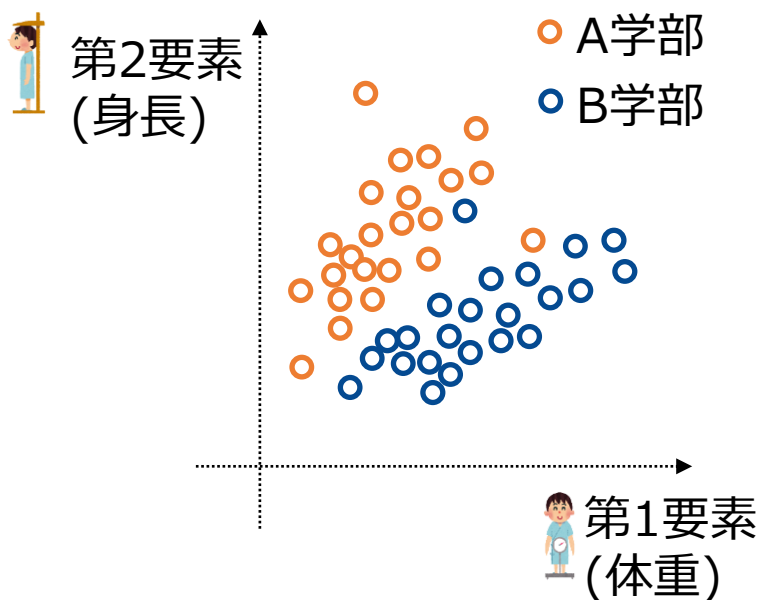
- データを理解するための方法
- 直感的な理解を支援
- 様々な可視化手法があり，適切に利用すること重要

ID	属性1	属性2	属性3	属性4	...	属性N
0	0.60	0.39	0.54	0.79	...	0.34
1	0.15	0.14	0.04	0.01	...	0.23
2	0.42	0.71	0.62	0.39	...	0.66
3	0.35	0.59	0.30	0.12	...	0.31
4	0.05	0.23	0.53	0.33	...	0.82
5	0.88	0.16	0.35	0.48	...	0.91
6	0.40	0.80	0.84	0.65	...	0.94
7	0.59	0.73	0.04	0.34	...	0.22
8	0.61	0.56	0.06	0.52	...	0.00
9	0.55	0.70	0.70	0.64	...	0.32
⋮	⋮	⋮	⋮	⋮	⋮	⋮



可視化からデータ分析へ

- データを見ることで，次に行う分析の方針決定につながる



- A学部とB学部のデータの広がり方はどのくらい違うの？
- 身長の増加と体重の増加に相関はあるの？

可視化と数学の意外(?)な関係

- 可視化，特に要素数を減らす方法の完全理解には，数学も必要
 - 本講義では詳しく述べませんでしたが…
 - 興味ある人は，「線形代数」(=足し算掛け算の深い世界)を勉強してみよう
- さらに，可視化の後の分析にも数学が必要

