

データサイエンス概論I & II データサイエンス総論I & II

可視化

九州大学 数理・データサイエンス教育研究センター

可視化とは？

数字だけ並べてもよくわからない → 図にするとよくわかる！

可視化の一般的な目的 (1/2)

- データに潜む性質を「目で見て」確認
- 例

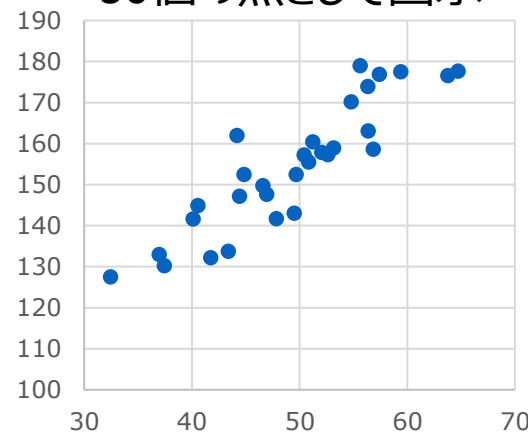


30人分の（体重，身長）データ

(49.5, 143.0)	(63.8, 176.6)	(56.4, 163.1)	(64.7, 177.7)	(44.9, 152.5)	(40.1, 141.6)
(46.6, 149.7)	(50.8, 155.6)	(52.1, 157.9)	(56.3, 173.9)	(41.8, 132.3)	(55.6, 179.0)
(56.9, 158.7)	(40.6, 144.9)	(57.4, 176.9)	(54.8, 170.2)	(53.2, 159.0)	(59.4, 177.5)
(47.8, 141.7)	(43.4, 133.8)	(37.5, 130.3)	(44.4, 147.2)	(49.7, 152.5)	(44.2, 162.0)
(51.2, 160.5)	(52.6, 157.3)	(32.5, 127.5)	(37.0, 133.0)	(46.9, 147.7)	(50.4, 157.3)



30個の点として図示

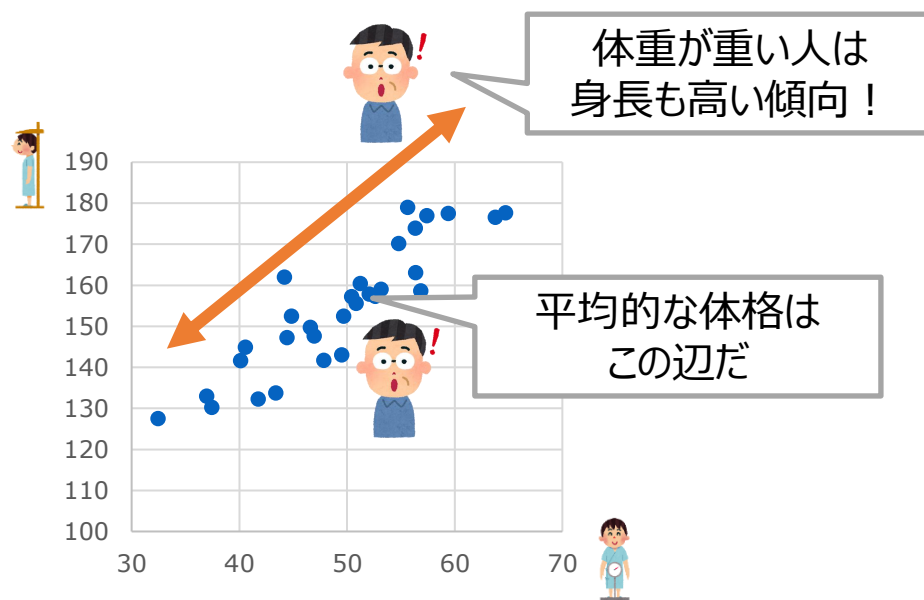


数字の羅列じゃ
よくわからない



可視化の一般的な目的 (2/2)

- 数値で直接見てもわかりにくいデータの性質を発見できる
 - データ全体の傾向 (分布や相関)
 - 個々のデータの性質 (平均に近いデータ, 珍しいデータ, 異常データ)

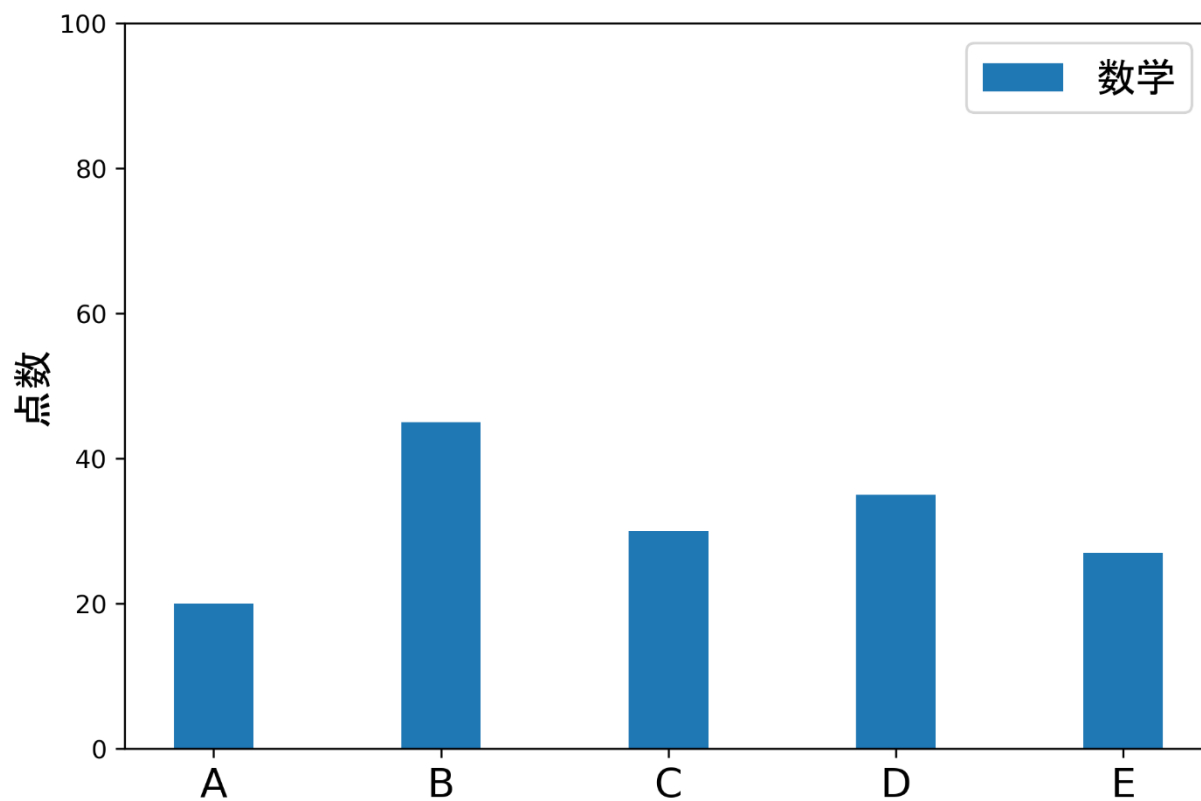


- 可視化法によっては, 変化 (動き) や関係も表現可能

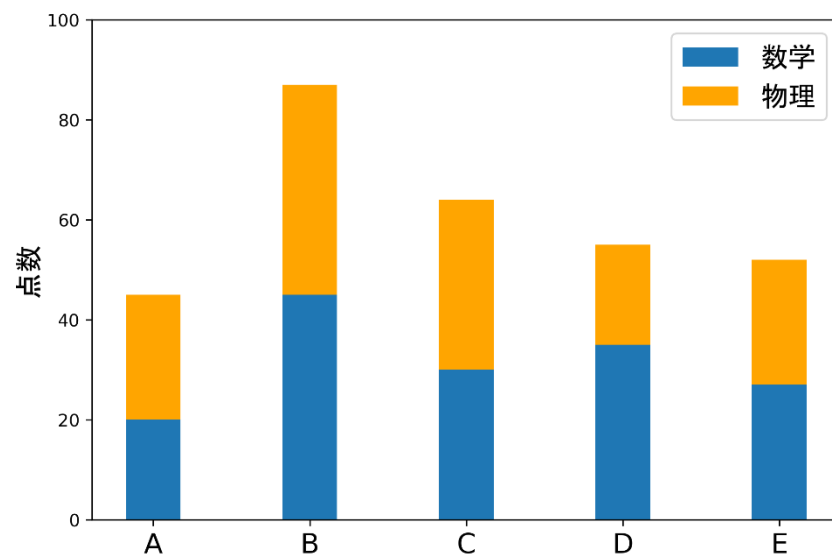
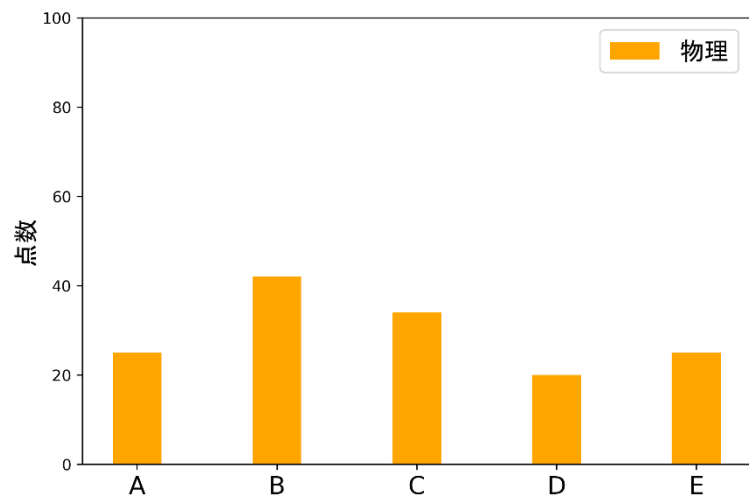
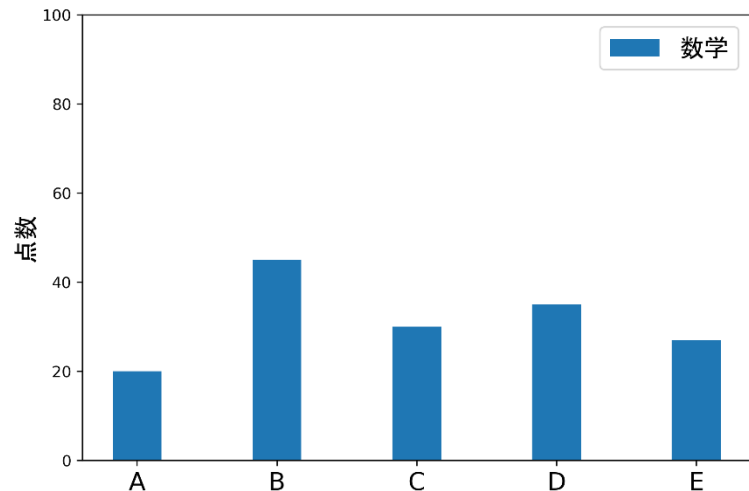
基本的な可視化

ここはサラッと

棒グラフ(bar chart)

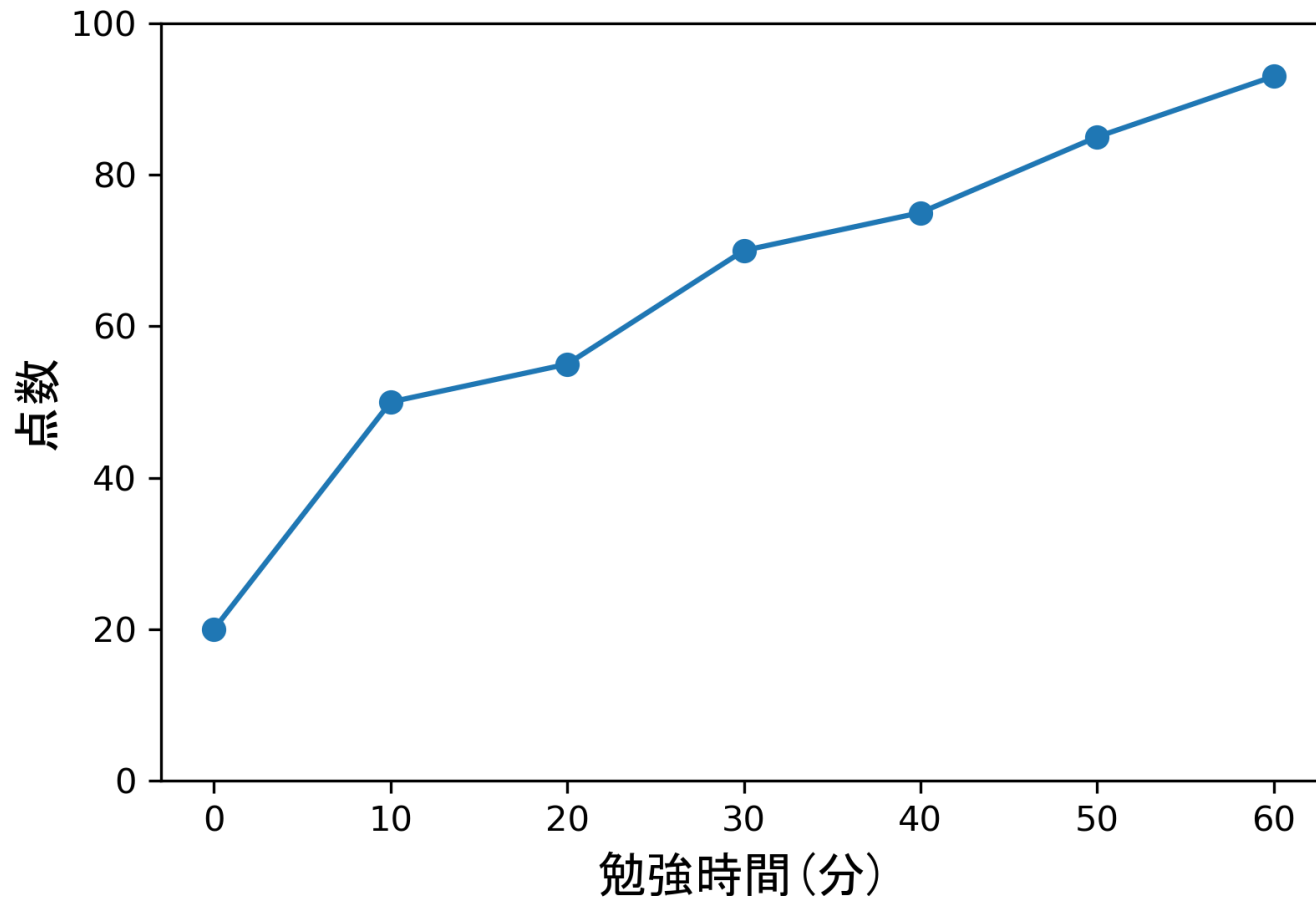


積み上げ棒グラフ

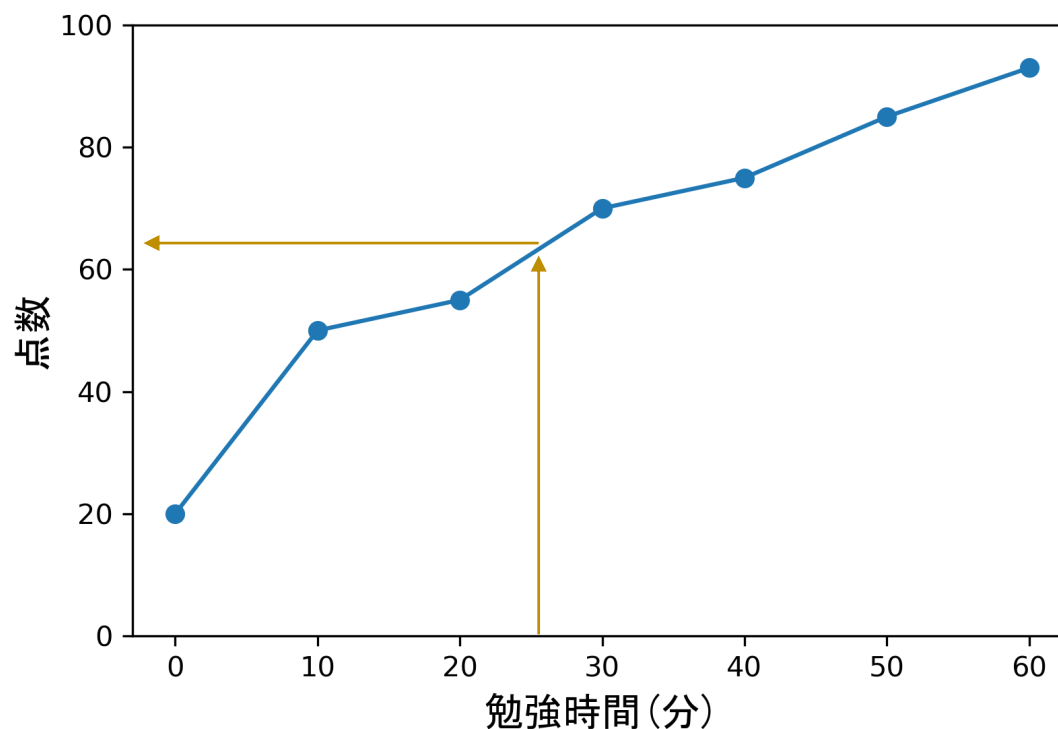


積み上げ棒グラフ

折れ線グラフ (line graph)



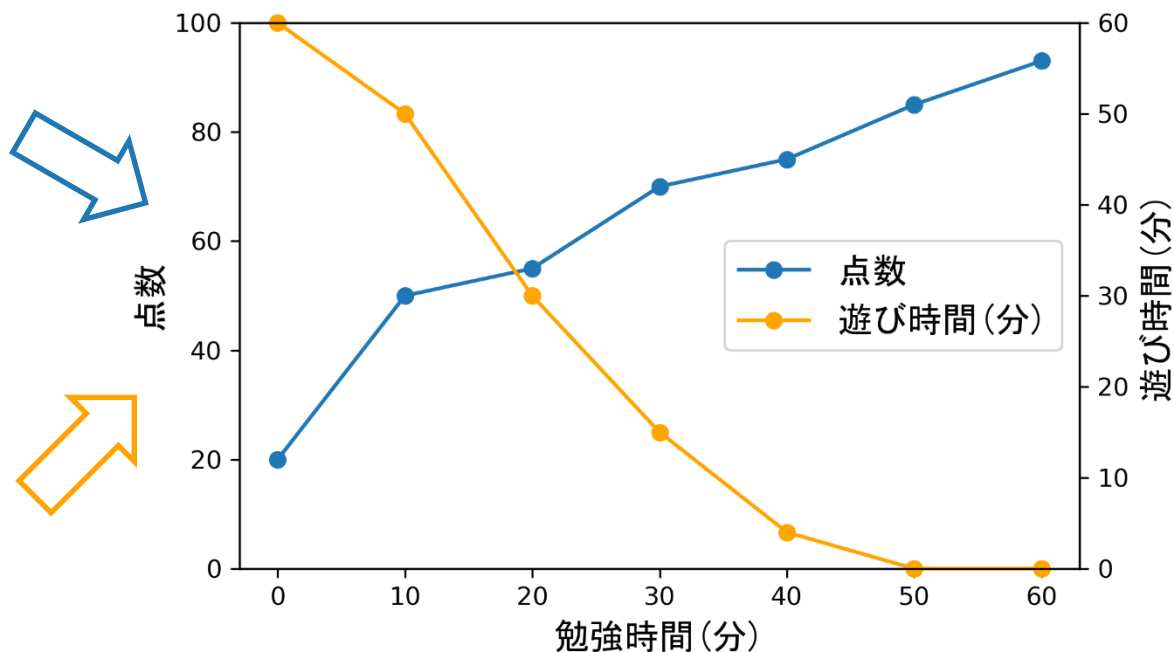
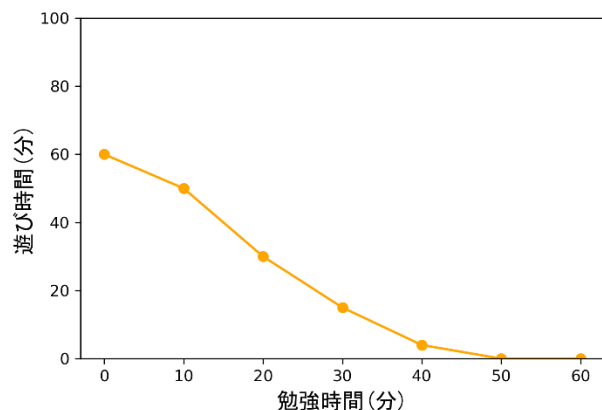
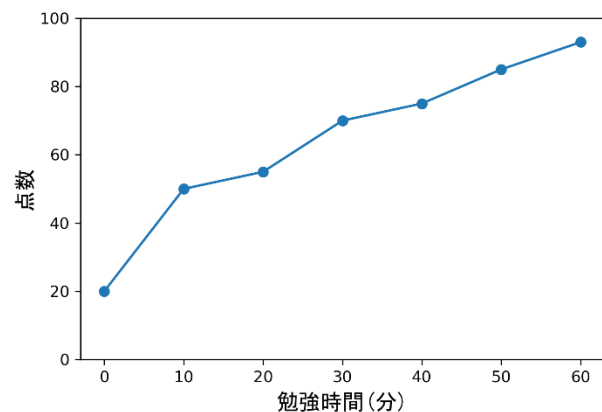
(前に言いましたが) 折れ線グラフでは、なぜ点を線で結ぶのか？



- 計測していないデータを「予測」
- なので横軸に「りんご みかん もも…」のような質的データは使えない

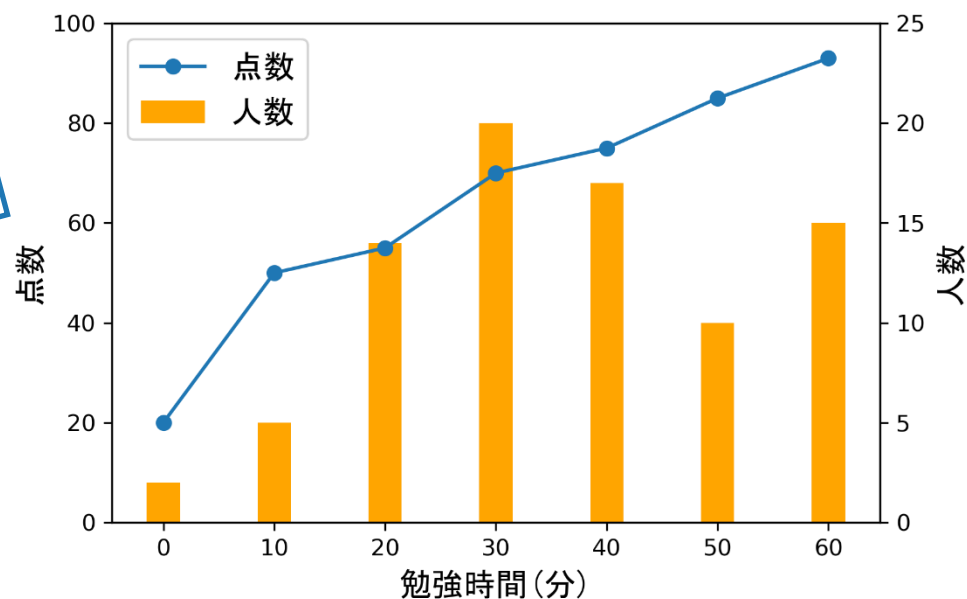
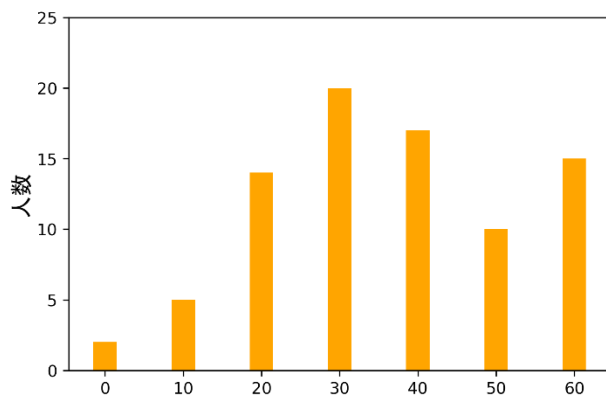
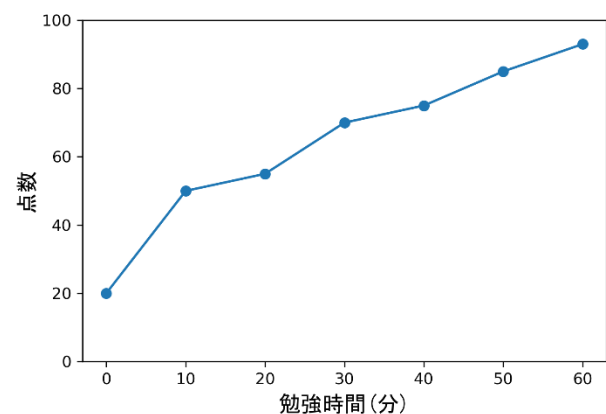
二軸グラフ：

横軸は同じで縦軸が異なる2つのデータの折れ線グラフを重ねたもの



- コンパクトになるだけでなく、両データの関係もわかる
 - 例：点数が増えると、遊び時間は減る

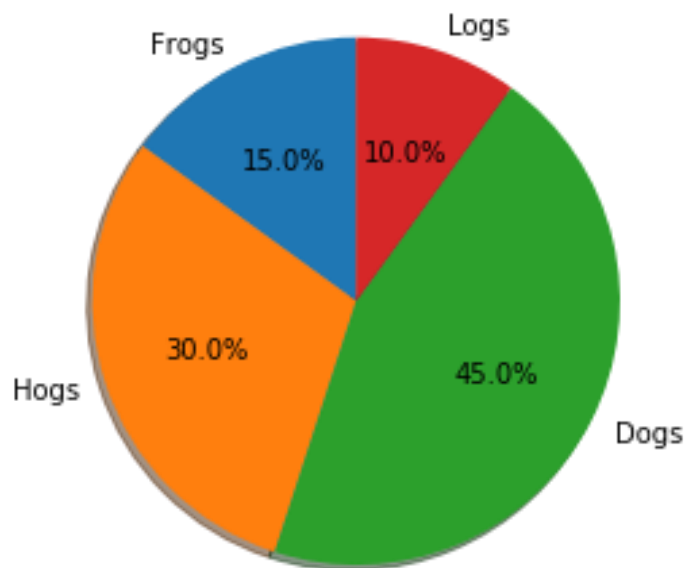
複合グラフ： 棒グラフ＋折れ線グラフ



- 二軸グラフと同様，コンパクトで，さらに両データの関係もわかる

円グラフ (pie chart)

- 割合を表現するには好適



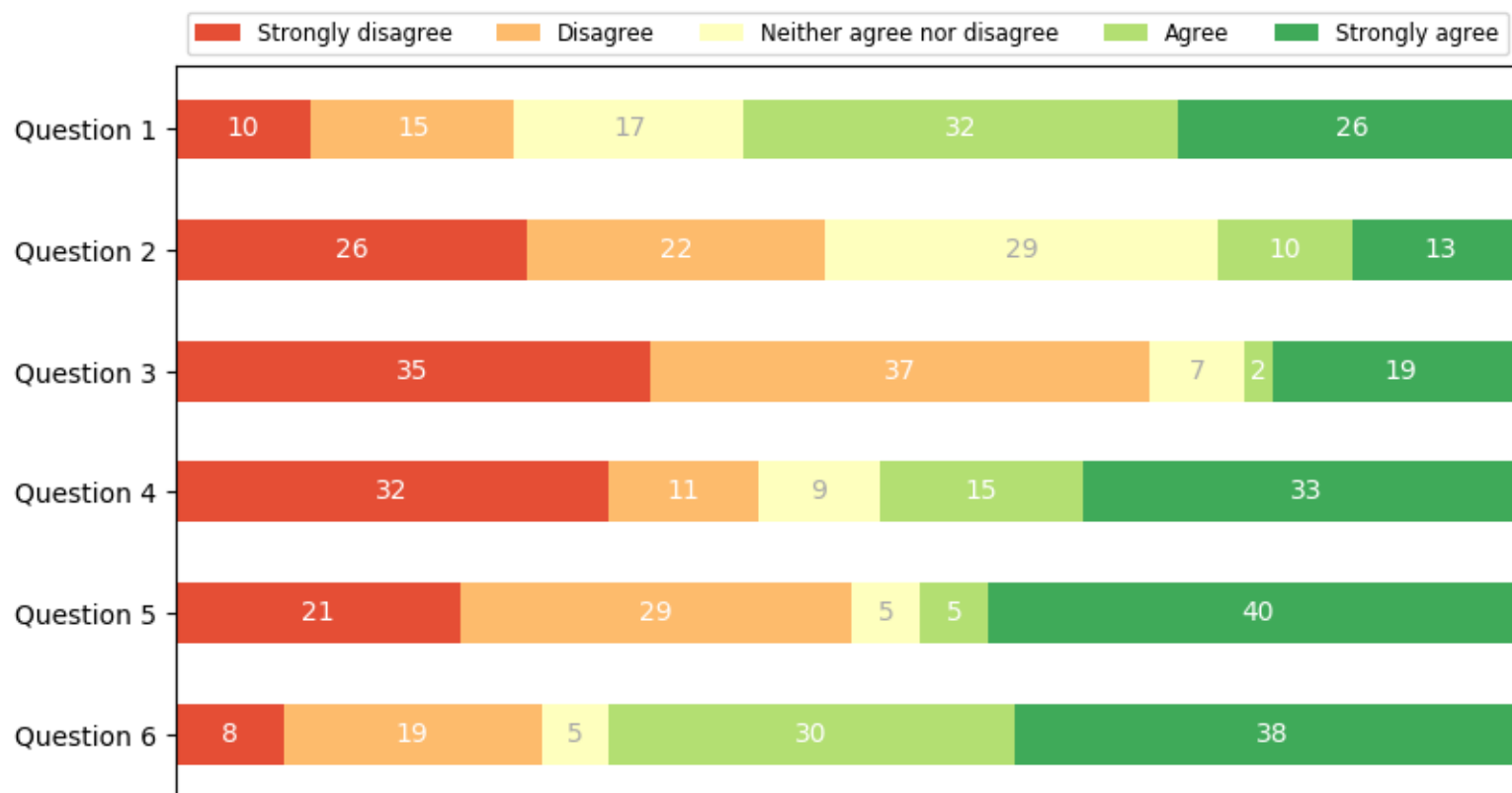
https://matplotlib.org/3.1.1/gallery/pie_and_polar_charts/pie_features.html#sphx-glr-gallery-pie-and-polar-charts-pie-features-py

- 複数の円グラフを, ネスト表現することも可能



https://matplotlib.org/3.1.1/gallery/pie_and_polar_charts/nested_pie.html#sphx-glr-gallery-pie-and-polar-charts-nested-pie-py

棒グラフを使っても、 円グラフと同じような表現は可能



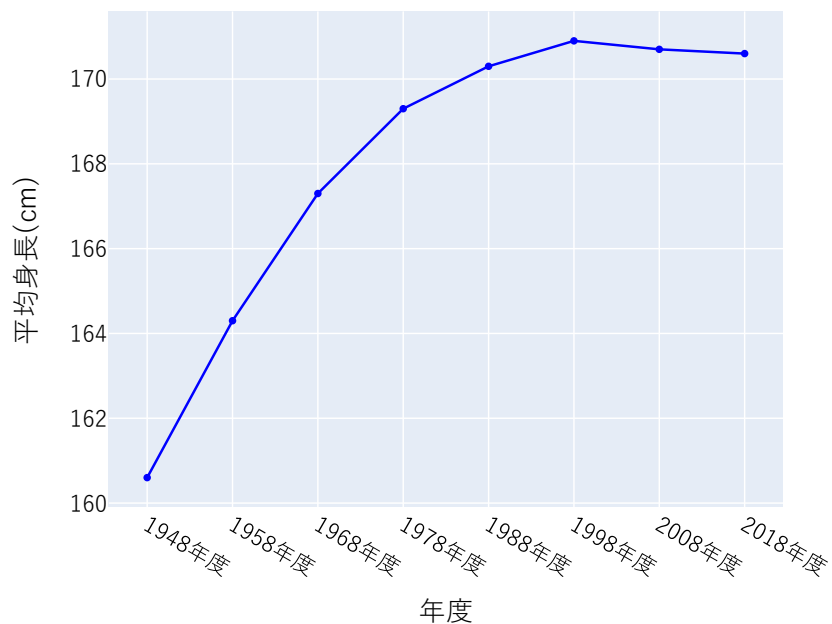
https://matplotlib.org/3.1.1/gallery/lines_bars_and_markers/horizontal_barchart_distribution.html#sphx-glr-gallery-lines-bars-and-markers-horizontal-barchart-distribution-py

メディアにだまされるな！ 折れ線グラフの罠(1/2)

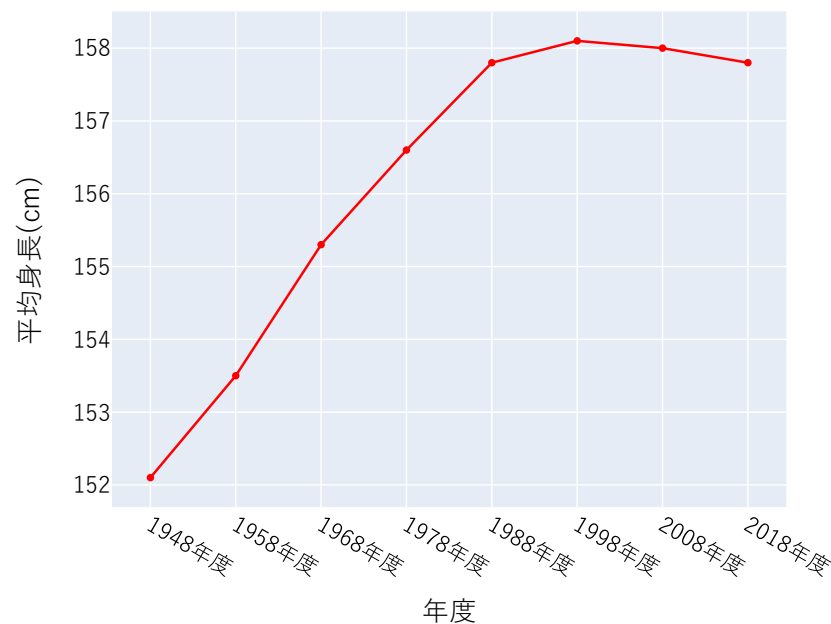


- 折れ線グラフを比較するときは**縦軸**に要注意！

男性17歳の平均身長推移



女性17歳の平均身長推移

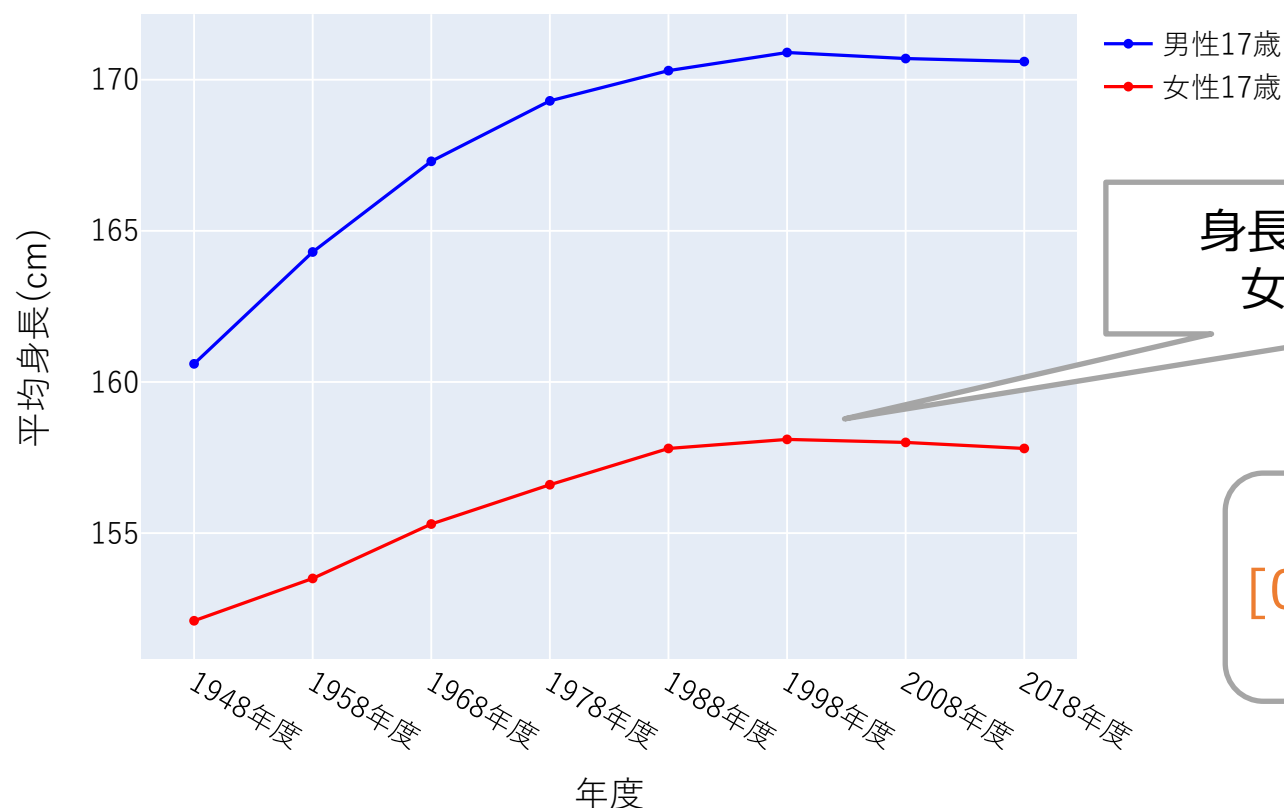


一見すると、男性女性で同じ変化に見えるが、**縦軸の範囲や幅が異なる！！**

メディアにだまされるな！ 折れ線グラフの罠(2/2)

- 同じグラフ描画内に描画するとよい

平均身長推移



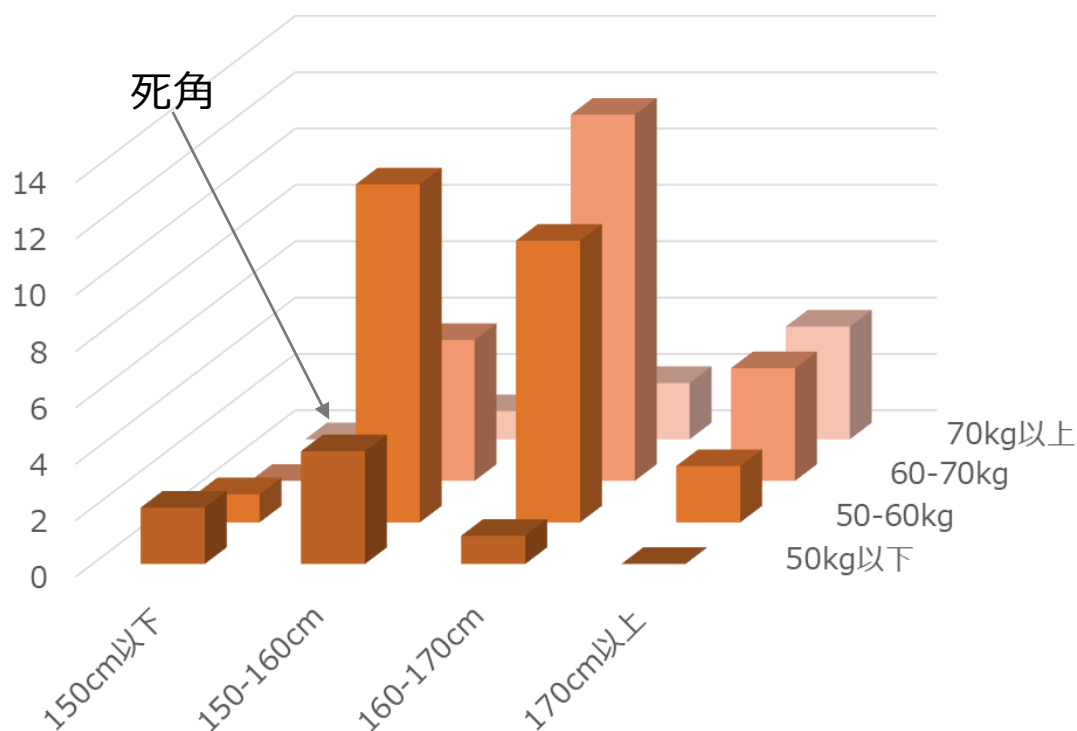
身長増加率・増加量は
女性のほうが小さい

ちなみにこのグラフ,
[0,180]の範囲で描くと
全く違う印象に!

メディアにだまされるな！ 3D棒グラフの罠



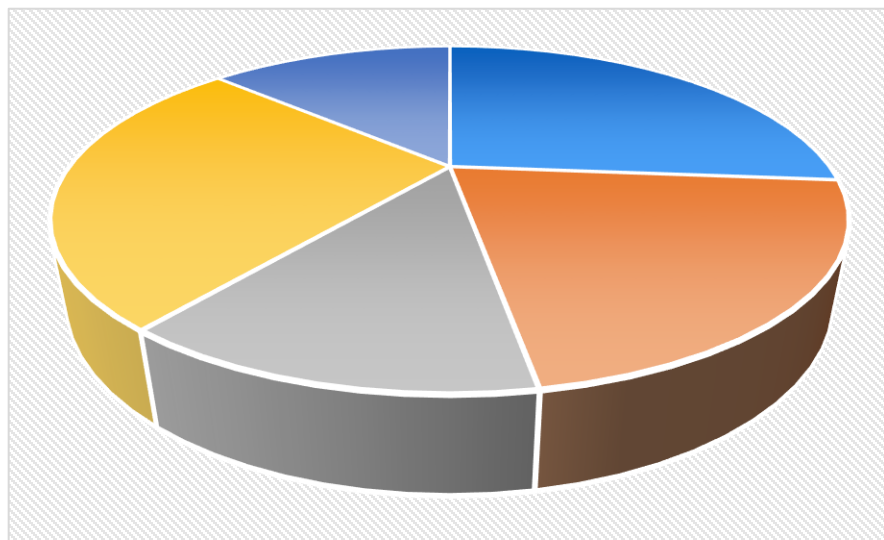
- 「死角」がある
- さらに高さを直接比較しにくい



メディアにだまされるな！ 3Dパイチャートの罠

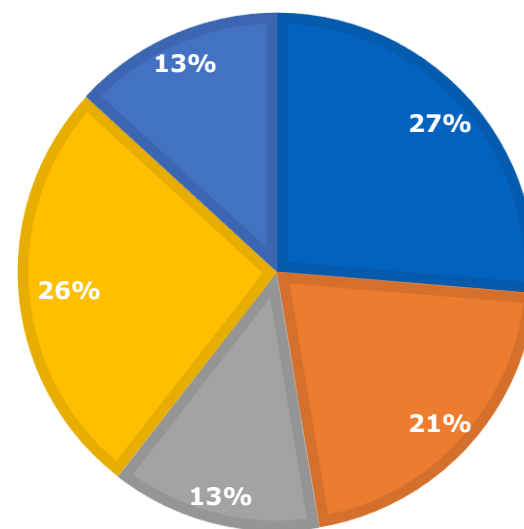


- 立体的になることで、面積の大小関係が異なるように見える
 - 見栄えはいいが、実際の値がなければ、誤解してしまう



灰色 > 水色に見える！！

=



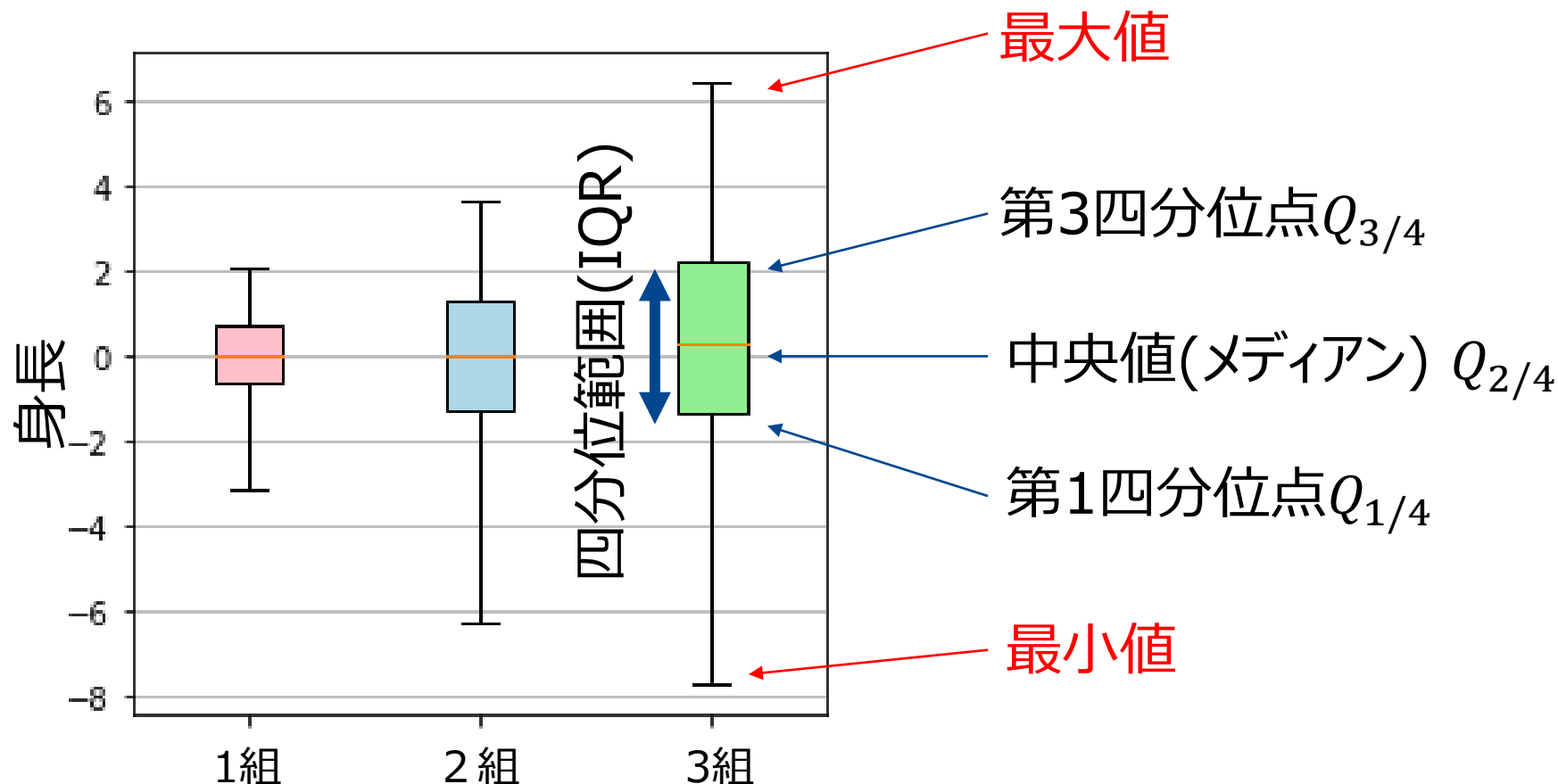
ホントは灰色 = 水色！！

- 使うのは避けたほうが吉

箱ひげ図

このへんもサラッと

箱ひげ図(box plot) : データ集合の広がりを視覚化



参考 : https://matplotlib.org/gallery/statistics/boxplot_color.html#sphx-glr-gallery-statistics-boxplot-color-py

四分位数？

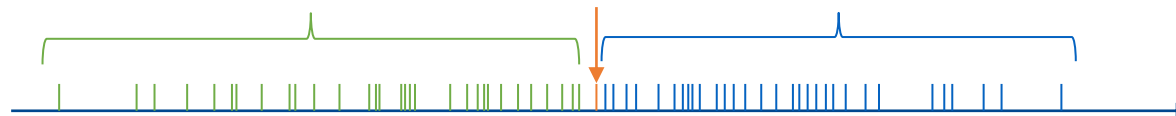
63人の身長分布



中央値
(メディアン)

下位31人

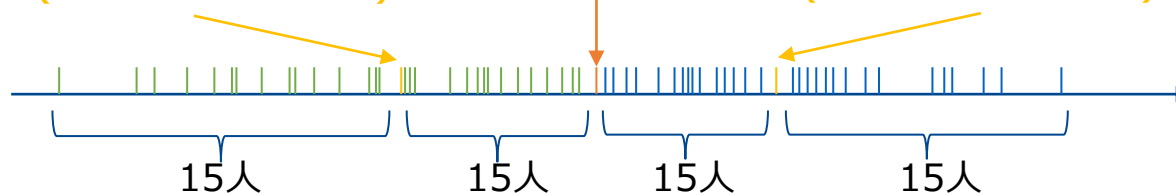
上位31人



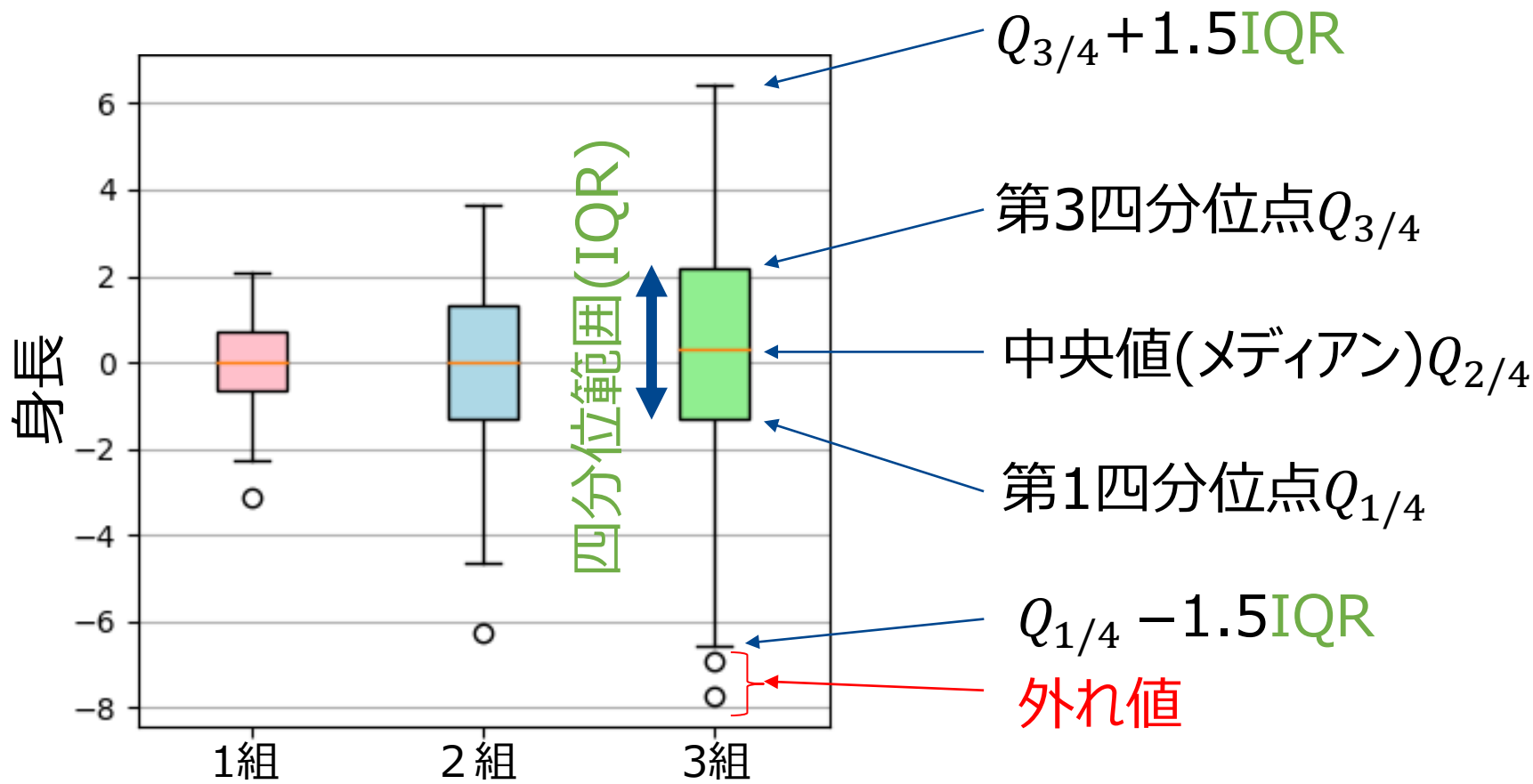
第1 四分位数
(25パーセンタイル)

中央値
(メディアン)

第3 四分位数
(75パーセンタイル)



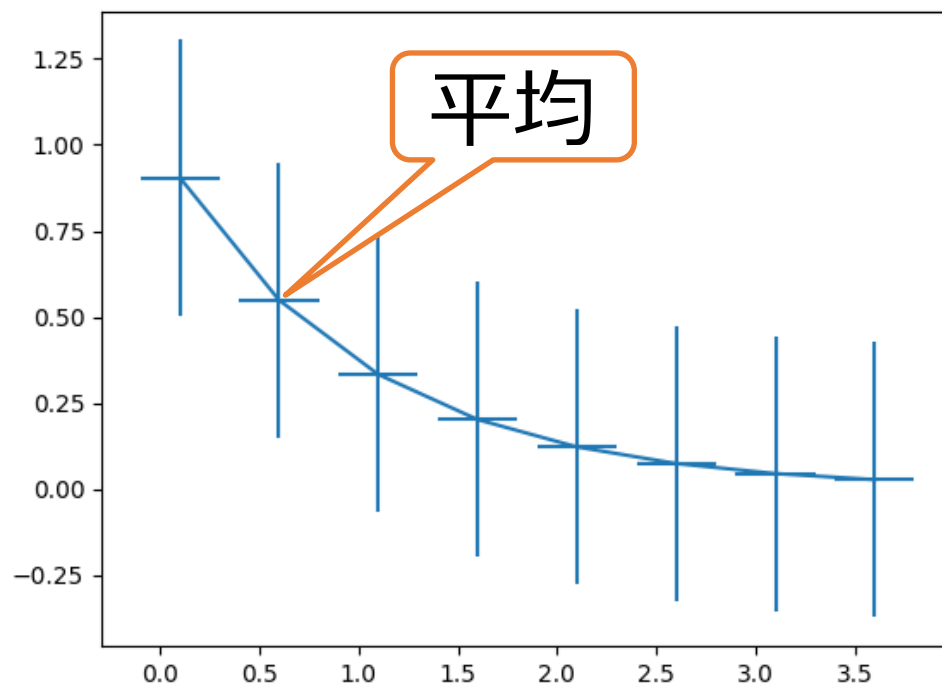
箱ひげ図(box plot) : こういうバージョンもあります



参考 : https://matplotlib.org/gallery/statistics/boxplot_color.html#sphx-glr-gallery-statistics-boxplot-color-py

箱ひげ図の親戚① エラーバー

- 折れ線× (平均 + 範囲)



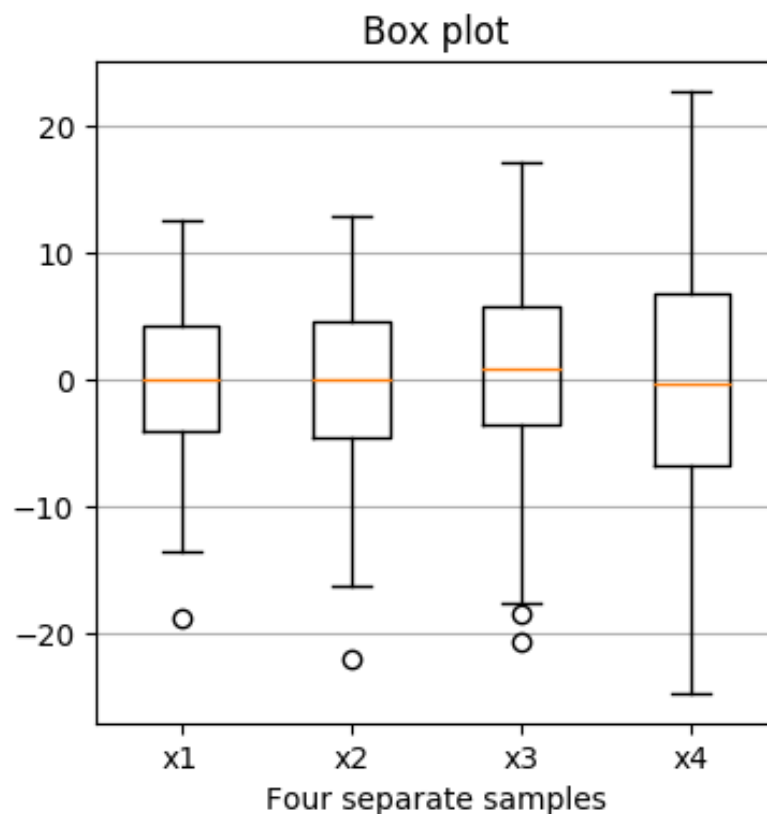
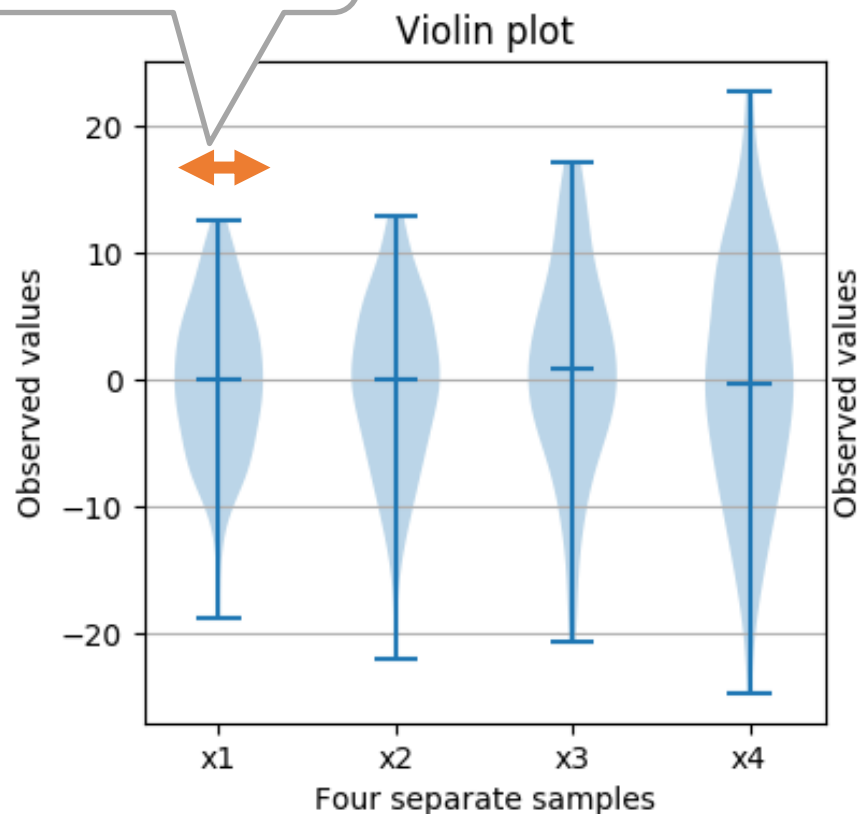
<https://matplotlib.org/3.1.1/gallery/statistics/errorbar.html#sphx-glr-gallery-statistics-errorbar-py>

- エラーバーの範囲
 - (最小値, 最大値)
 - (平均-標準偏差, 平均 + 標準偏差)
 - 信頼区間
 - ...

ということで, 色々ある → 何かを正しく明記する必要がある.

箱ひげ図の親戚② バイオリン図： 「出やすさ」も表現する

広がっているほど
その辺の値が出やすい



https://matplotlib.org/3.1.1/gallery/statistics/boxplot_vs_violin.html#sphx-glr-gallery-statistics-boxplot-vs-violin-py

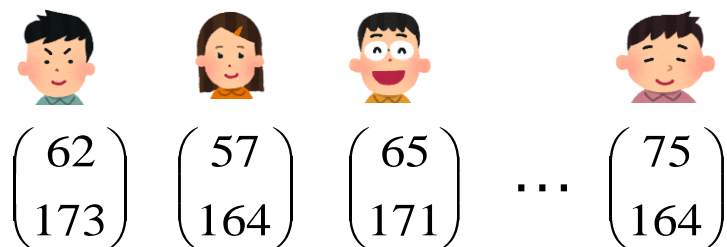
散布図

これまでも何度か出てきた **2次元データ** 集合の可視化.
非常によく使います

2次元データ集合の 散布図

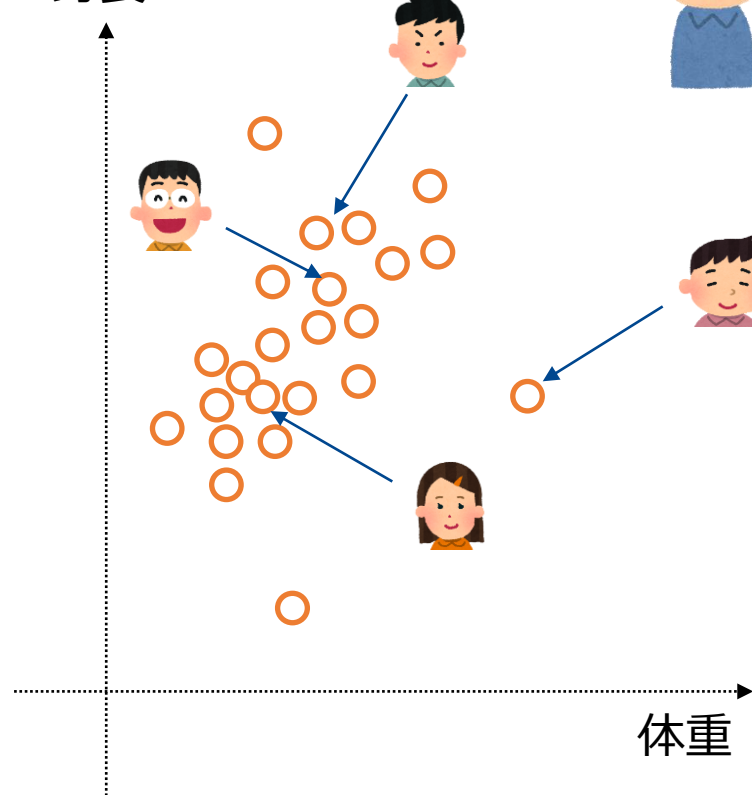
平均や
ばらつきが
一目瞭然

2次元データの集合



数字の羅列じゃ
よくわからない

身長

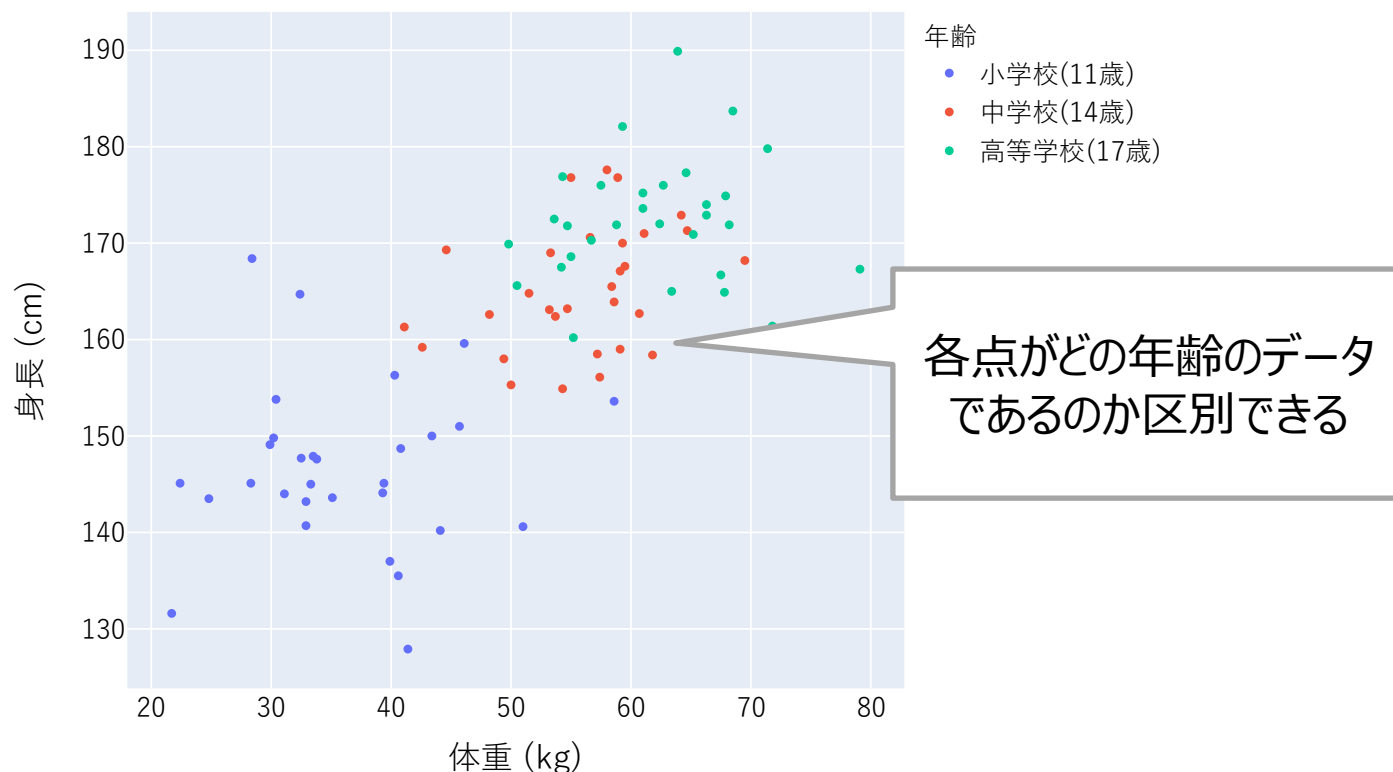


体重

散布図の拡張(1/2) : 色 + 散布図

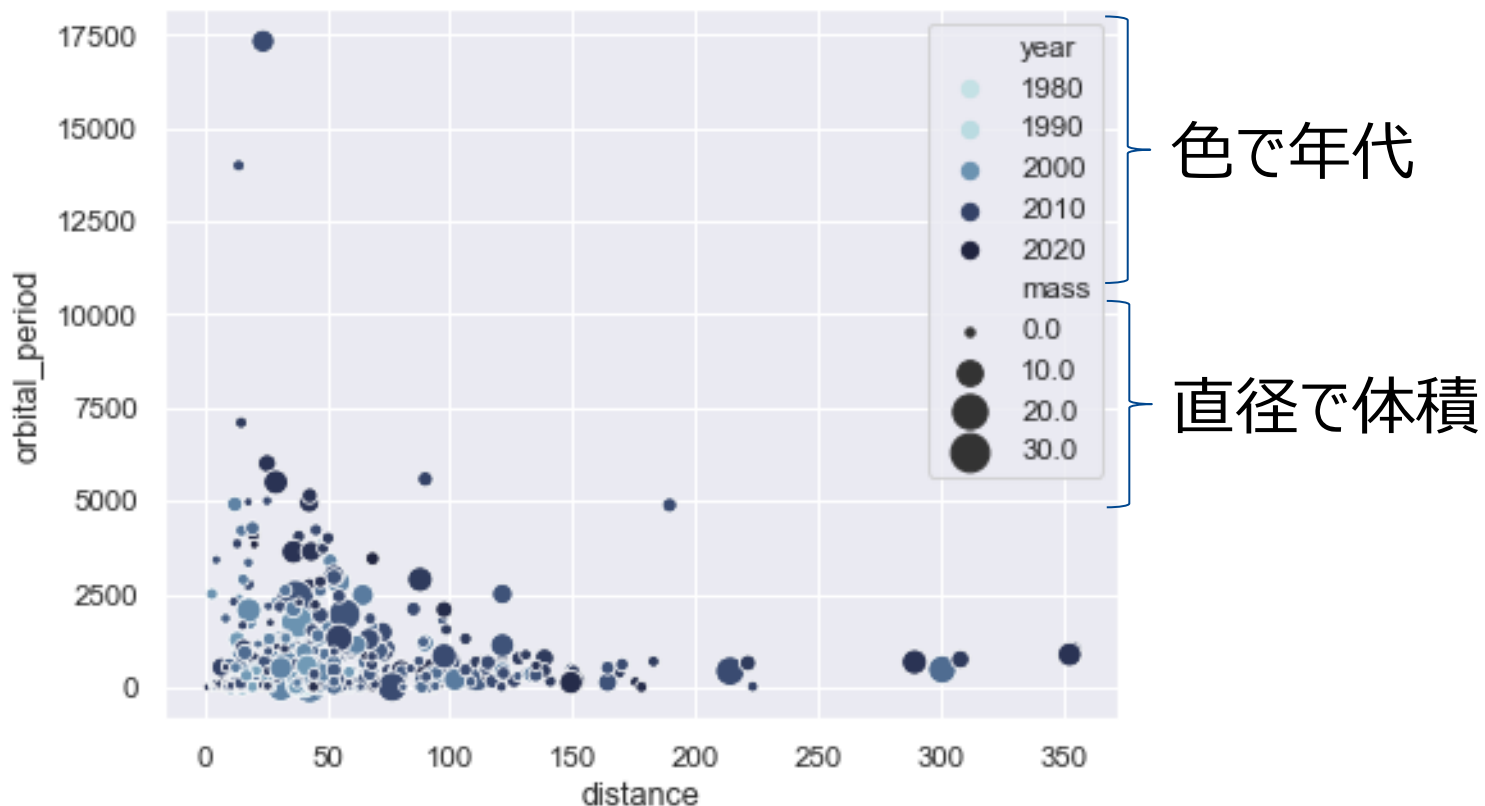
- 色によって3種類の値(=3次元データ) を可視化可能
 - 形や点の大きさを変えることでさらに多くの情報を同時に可視化できる

男性（最終学年）の体重と身長散布図



散布図の拡張(2/2) : 色 + 大きさ + 散布図

- 4次元データも「頑張れば」散布図で表現可能(見づらいのであまりお勧めしない)

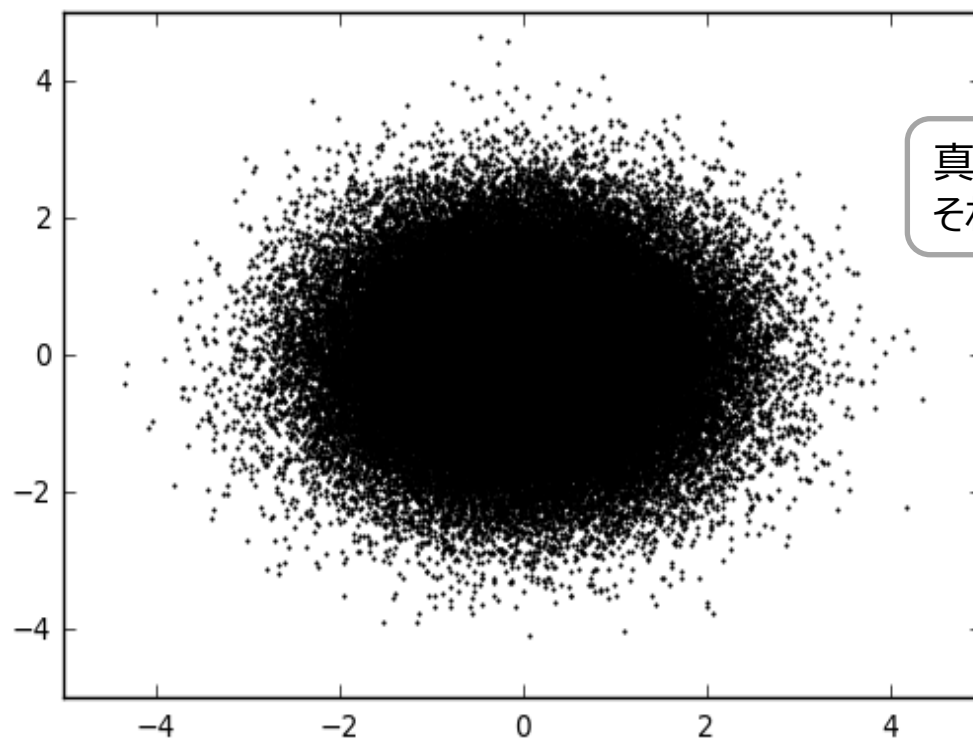


http://seaborn.pydata.org/examples/scatterplot_sizes.html (“planet”データ)

- さらに高次元のデータについては、さらに厳しい...

データ数が膨大になると不向き

- 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…



- 対処法→ヒストグラムやヒートマップ(後述)

散布図はデータ集合の可視化の強力ツールだが、 拡張も必要

散布図

比較的少数の2次元データ集合に適する

データが高次元に

平行座標プロット
散布図行列
主成分分析
多次元尺度構成法(付録)
tSNE(付録)

データが大量に

(2次元)ヒストグラム
ヒートマップ

高次元データの可視化

いよいよ可視化手法の本領発揮！？

平行座標プロット(parallel coordinates plots)

折れ線を利用して高次元データを表示(1/2)

(62, 173, 78, 26)



折れ線グラフのような
内挿目的ではい、
次元間での
つながりを表現

200

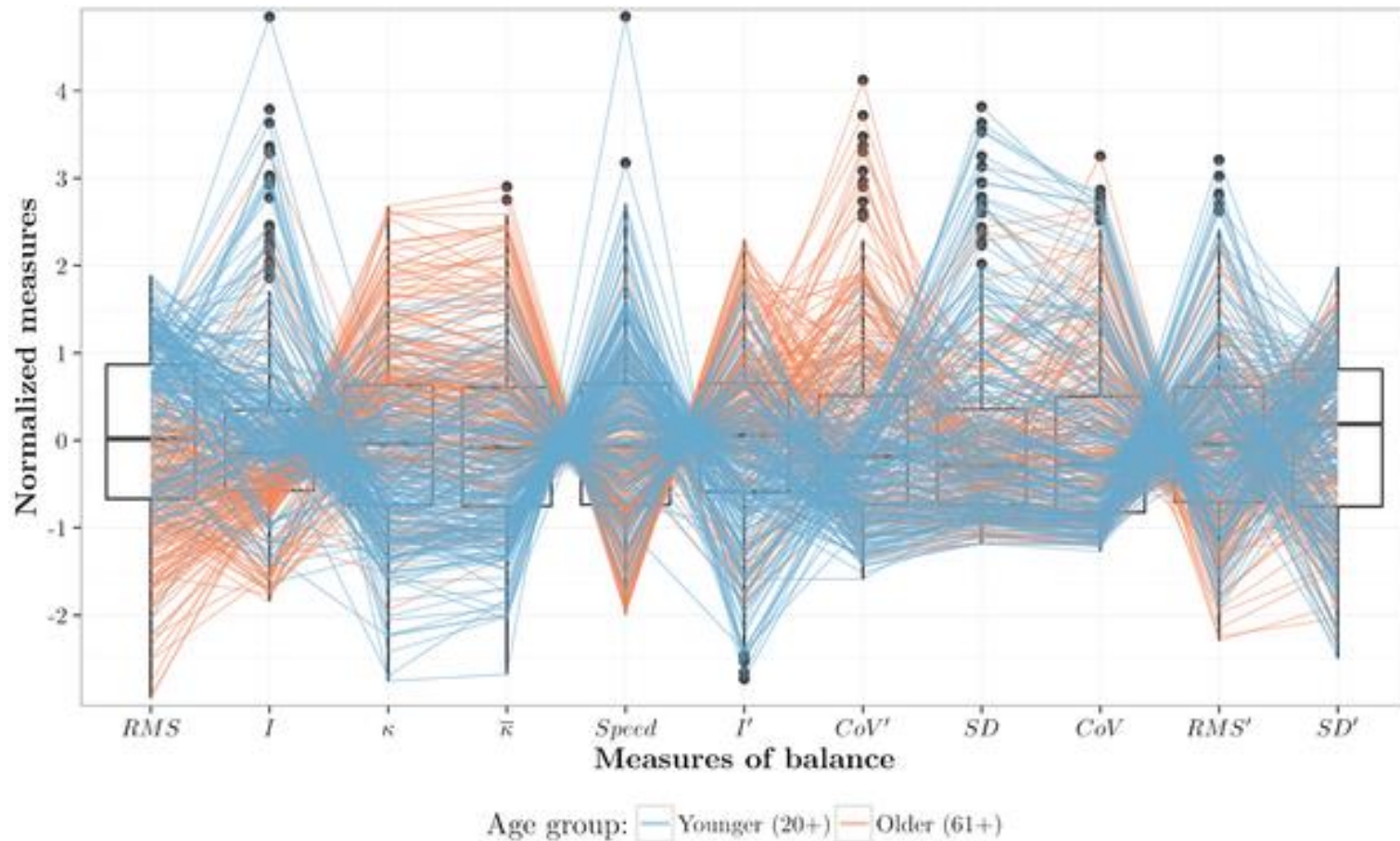
0



こんな調子で
何次元でもOK

平行座標プロット(parallel coordinates plots) 折れ線を利用して高次元データを表示(2/2)

- グループで色分けするとよい。ただデータが多すぎると見づらい

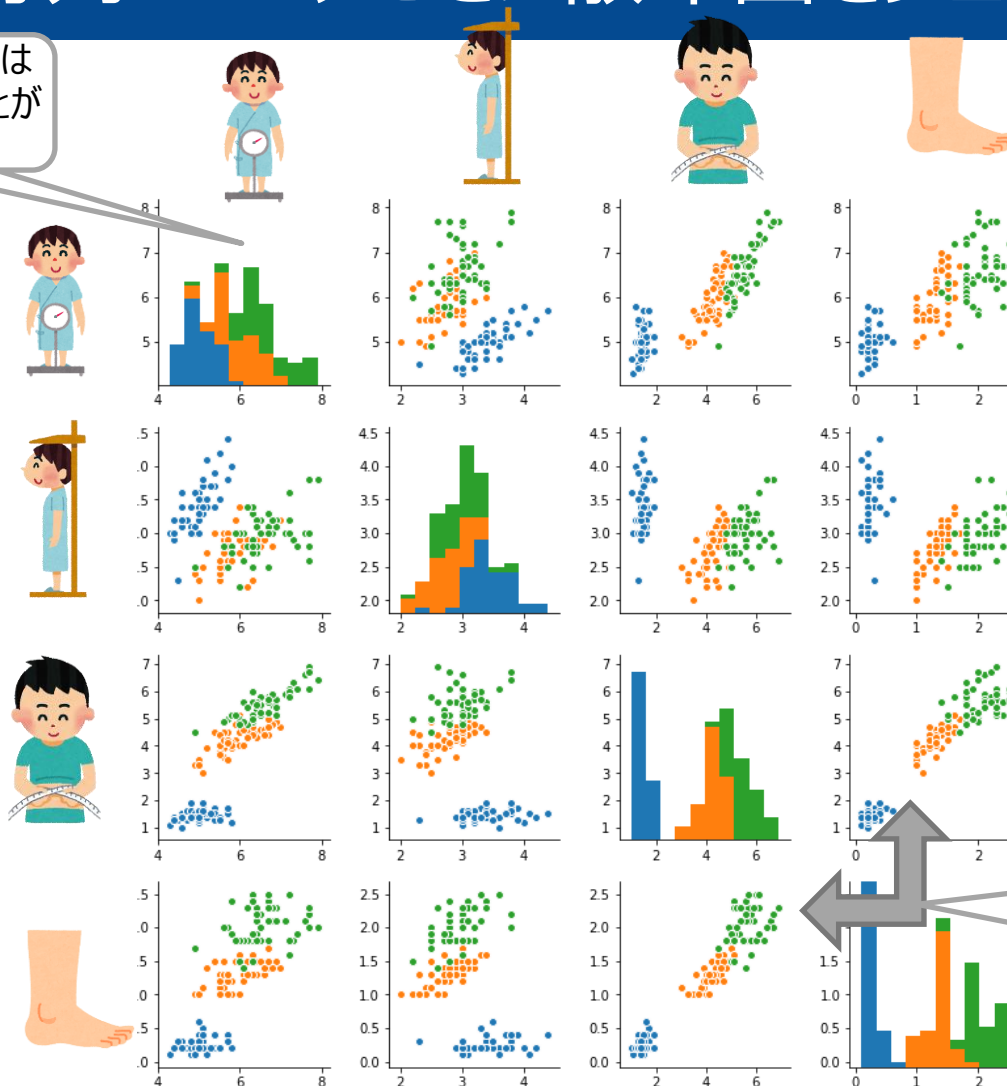


「2成分だけ見れば、2次元だ!」 散布図行列 – ペアごとに散布図をプロット

対角(斜め)成分には
ヒストグラムを書くことが
多い

(62, 173, 78, 26)

↑このタイプの4次元ベクトル
データがたくさんある場合..

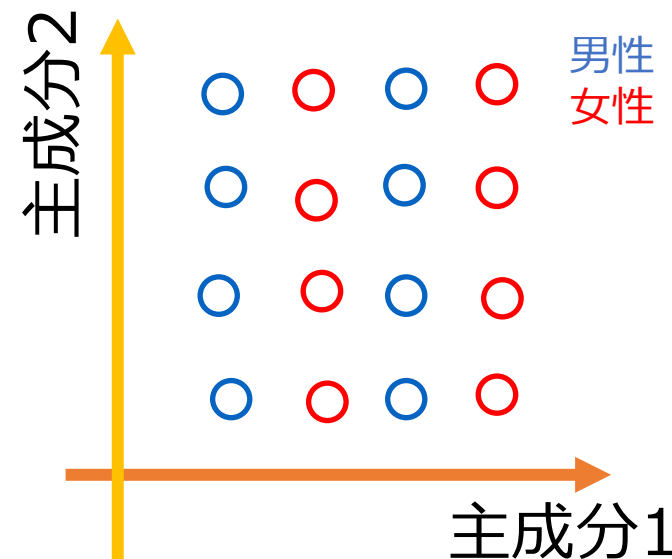
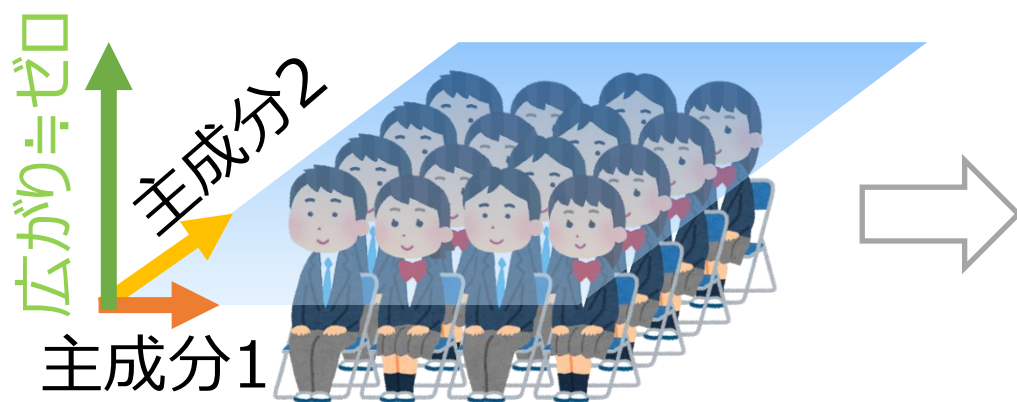


90度回転している
だけで、同じモノ

主成分分析(PCA)

上位2つの主成分を使えば散布図ができる!

- 「3次元→2次元」で以前使った例



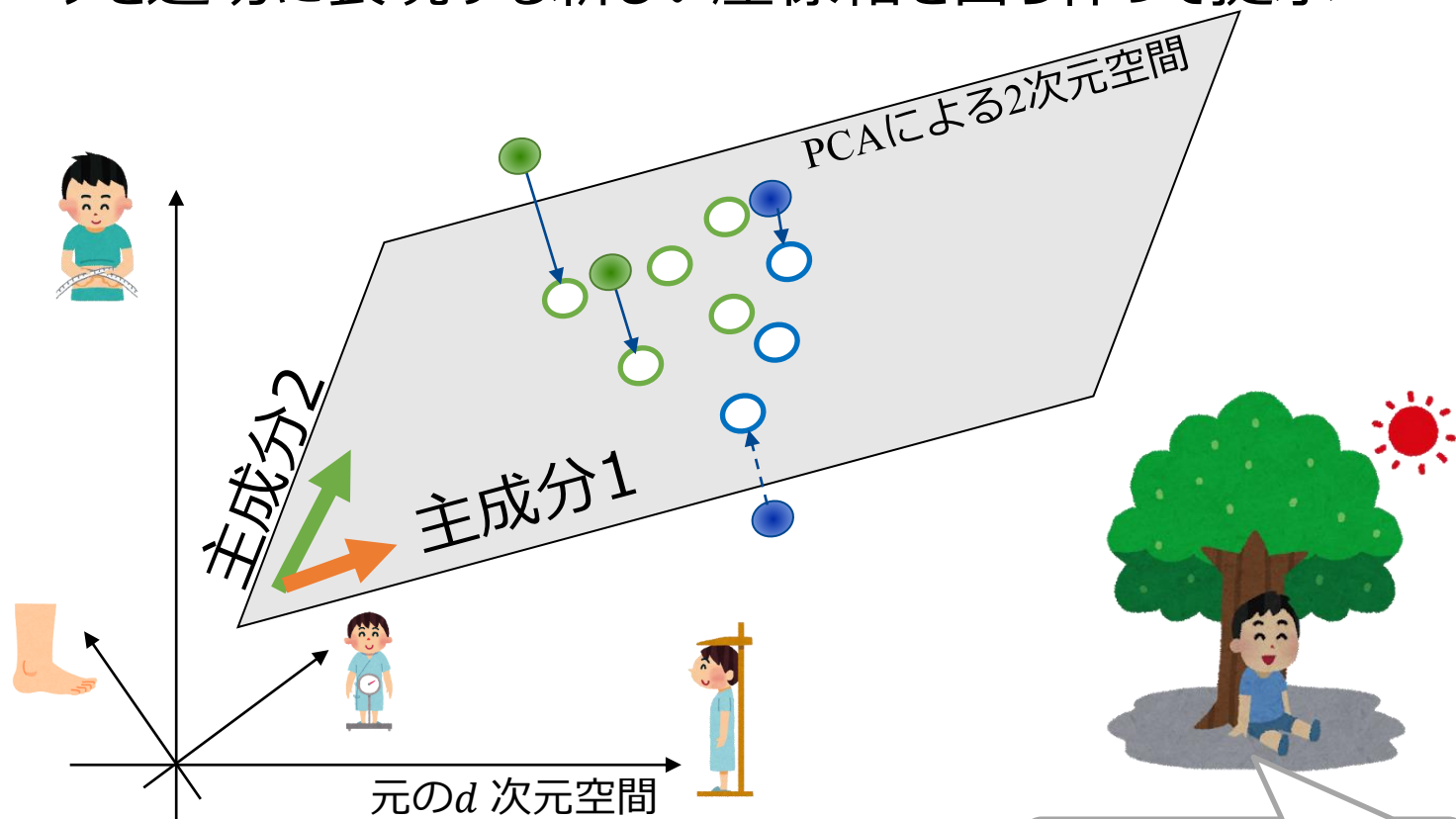
- ちなみに何次元データでも2次元(や3次元)にできます
 - 「線形代数に基づくデータ解析の基礎」主成分分析を参照

$$d\text{次元ベクトル } \boldsymbol{x} \rightarrow 2\text{次元ベクトル}(\boldsymbol{x} \cdot \boldsymbol{e}_1, \boldsymbol{x} \cdot \boldsymbol{e}_2)$$

第1主成分 \boldsymbol{e}_1 との内積
 第2主成分 \boldsymbol{e}_2 との内積

主成分分析(PCA)と散布図行列の違い

- 散布図行列：もともとある座標軸を二つ組み合わせる
- PCA：データを適切に表現する新しい座標軸を自ら作って提示

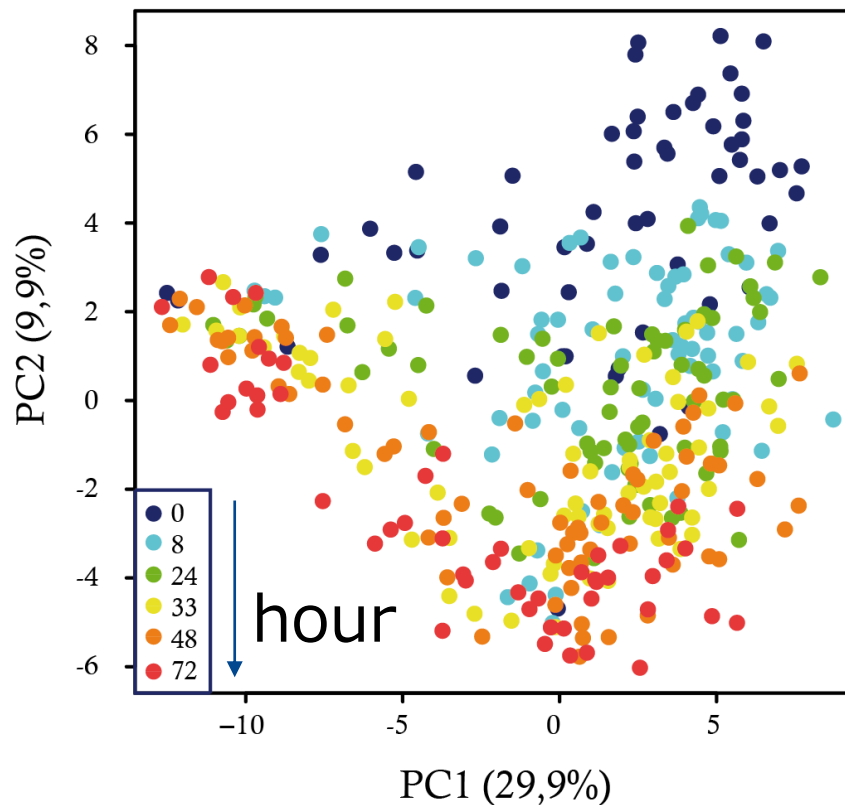


2次元(地面)に可視化した,
木の葉(3次元空間内に存在)の分布

主成分分析(PCA)

上位二つの主成分を求めれば...

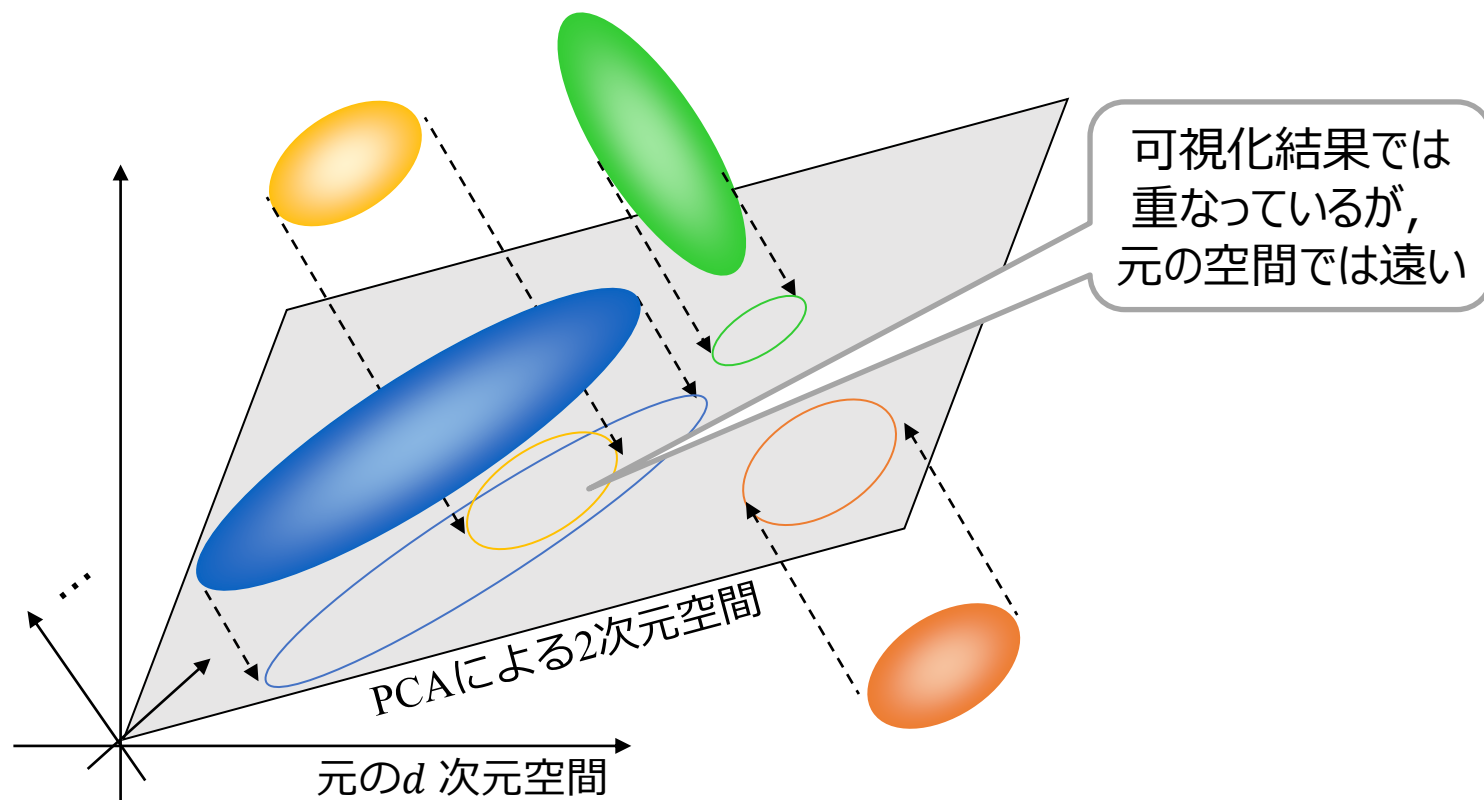
- 90次元の遺伝子発現データの時間的推移を2次元で表現



[Richard+, PLOS Biology, 2016]
<https://doi.org/10.1371/journal.pbio.1002585>

主成分分析(PCA)による可視化結果の見かた

- 可視化結果が重なっていない → 元の空間でも絶対重なっていない
- 可視化結果が重なっている → 元の空間で重なっているか不明

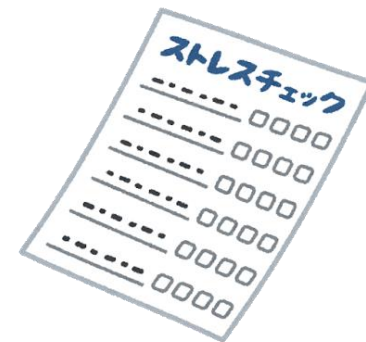
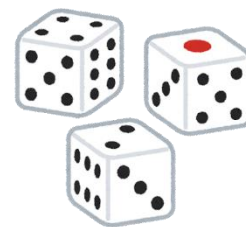


1次元データの頻度の可視化： ヒストグラム

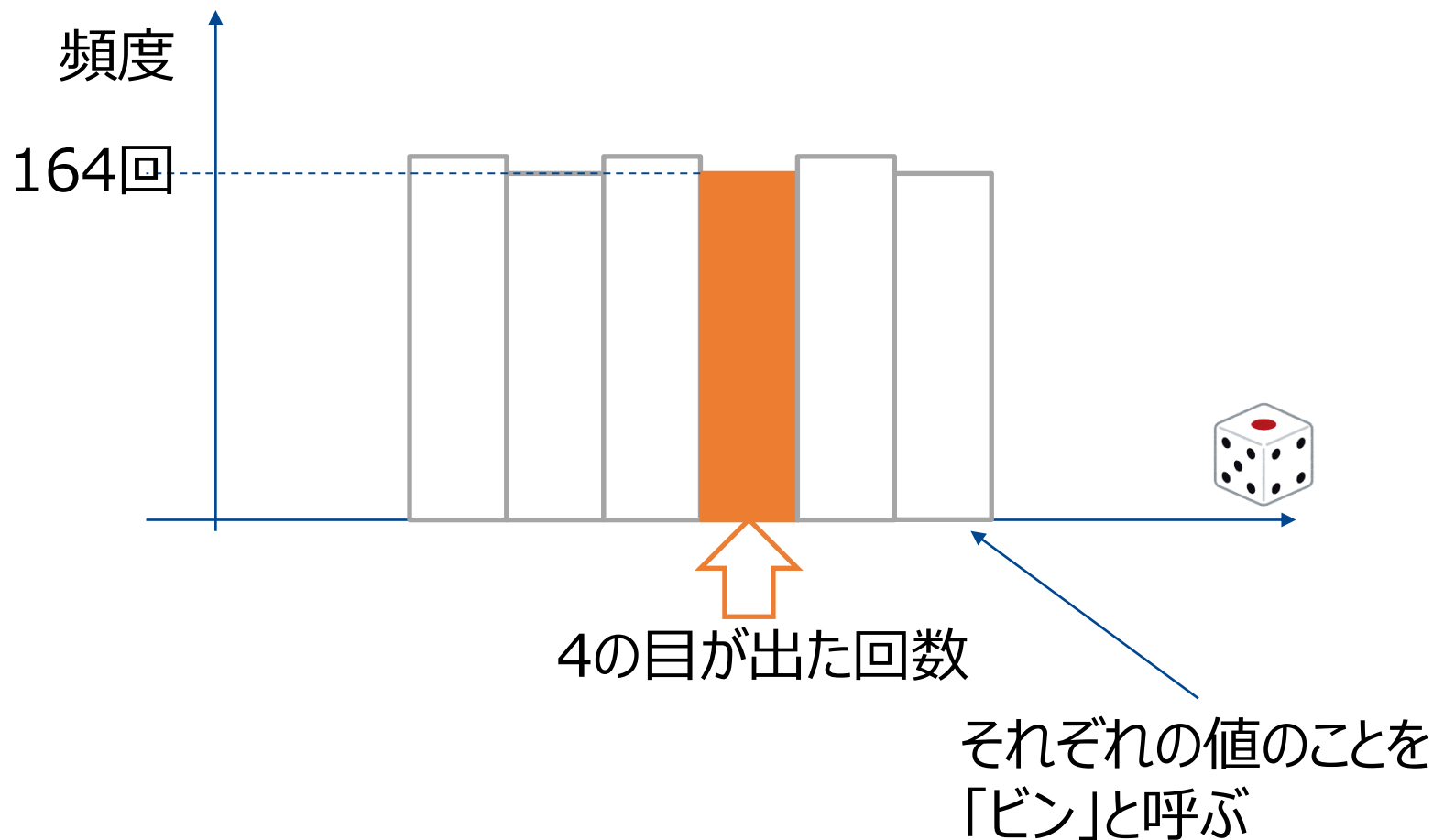
多くの人々の身長データをどう可視化するか？

頻度～わかりやすい場合

- さいころを1000回振って出た目の回数
 - 「1」が168回, 「2」が164回, ..., 「6」が164回
- 5段階アンケートの回答結果の集計
 - 「非常によい」が103名, 「よい」が30名, ..., 「非常に悪い」が0名
- 今日のメニュー注文者数
 - 「かつ丼」が58食, 「ラーメン」が102食, ..., 「高菜めし」が21食



ヒストグラムによる頻度分布の可視化: さいころを1000回振って出た目のヒストグラム



頻度～そのままでは計りにくい場合(1/2)

- あるクラスの学生の身長
 - ..., 167.301cmが0人, 167.302cmが1人, 167.303cmが0人, ...



- ピッタリ同じ身長の人にはほぼいないだろうから, 頻度は高々1

63人の身長の分布



ヒストグラムというより
(1次元)散布図みたい...

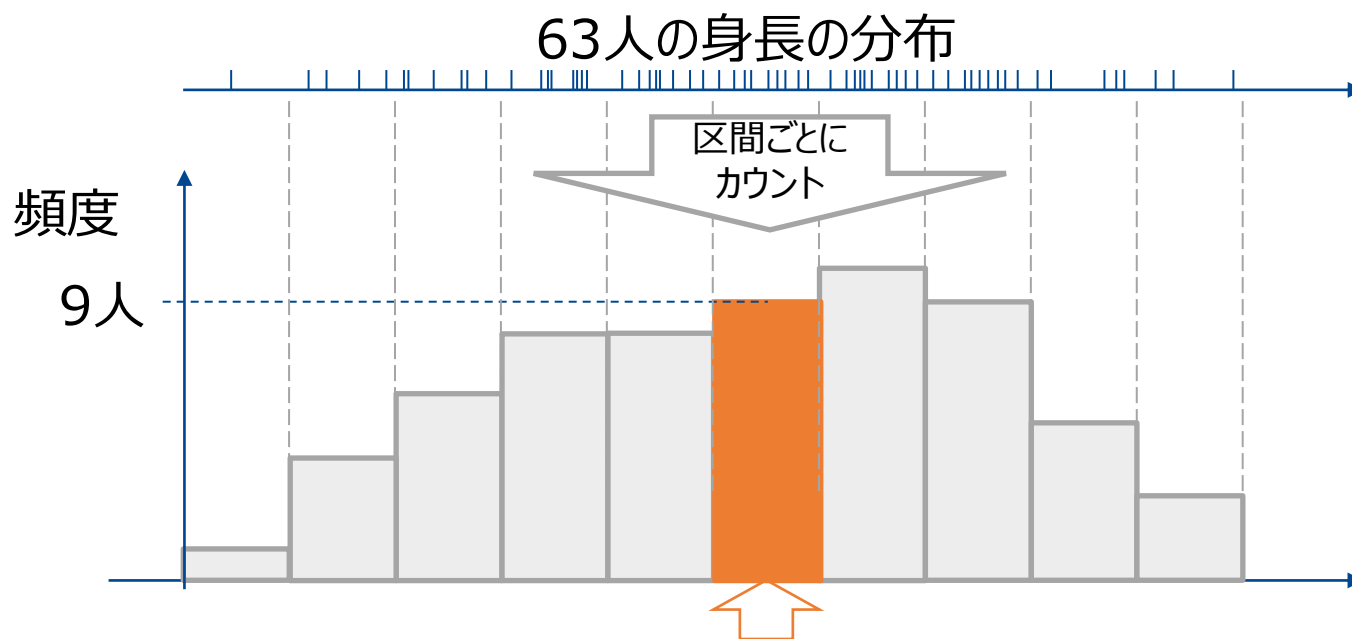


- 「値が連続的に変化する対象」の場合, 頻度は計りにくい

頻度～そのままでは計りにくい場合(2/2) → 区間を考えればOK

- あるクラスの学生の身長

- 「140cm未満」が1人, 「140-145cm」が2人, ..., 「160-165cm」が9人, ...



160cm~165cm ← それぞれの区間のことを「ビン」と呼ぶ

- 頻度がよくわかるように！

- ただし区間幅の設定によって集計結果が変わることに注意)

以上 2 ケースのまとめ

- データ x が取りうる値が**有限個**(例えば B 個)の場合
 - B 個のビンからなるヒストグラム
 - 1つのビン = 1つの値に対応
- データ x が取りうる値が**無限個**の場合
 - x の値の全範囲を B 個の区間に分ける
 - B 個のビンからなるヒストグラム
 - 1つのビン = 1つの区間に対応

= 値が連続的に
変化する場合

なんでもないような話ですが、これがあと(確率)の話で
重要なポイントになります

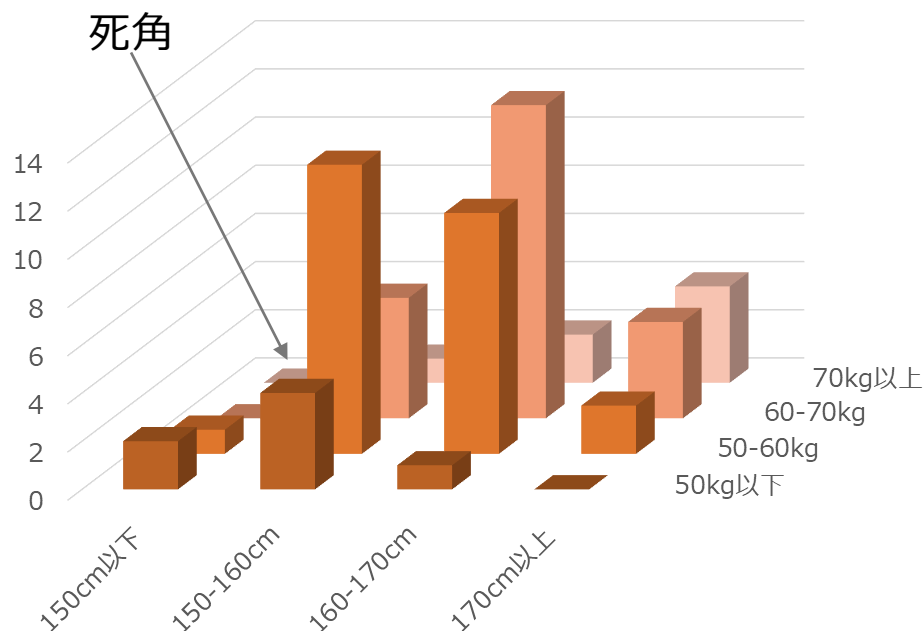
2次元データの頻度の可視化： 2次元ヒストグラムとヒートマップ

2次元ヒストグラム

● ヒートマップタイプ

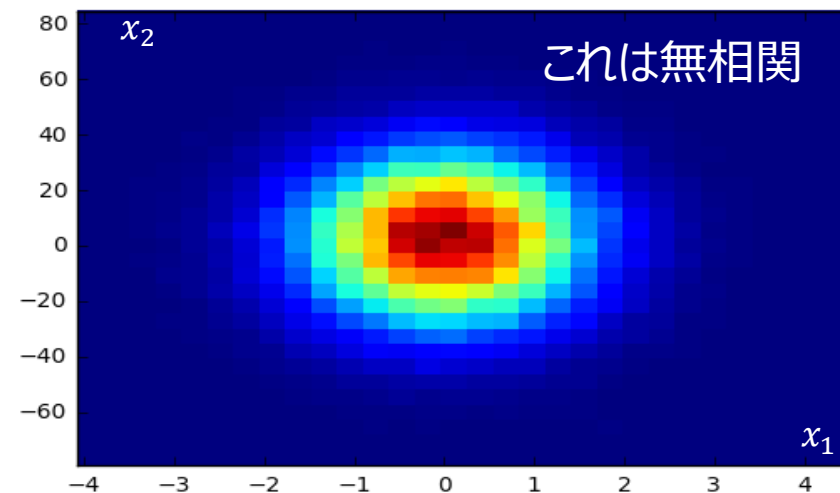
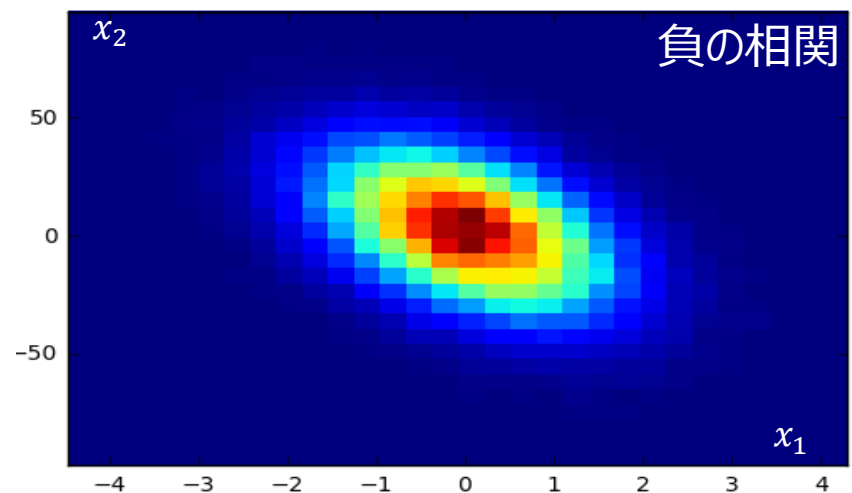
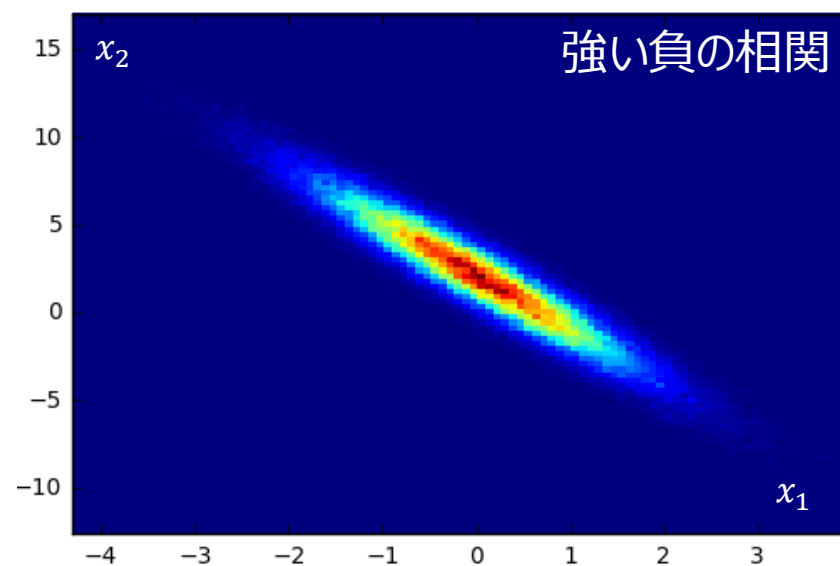
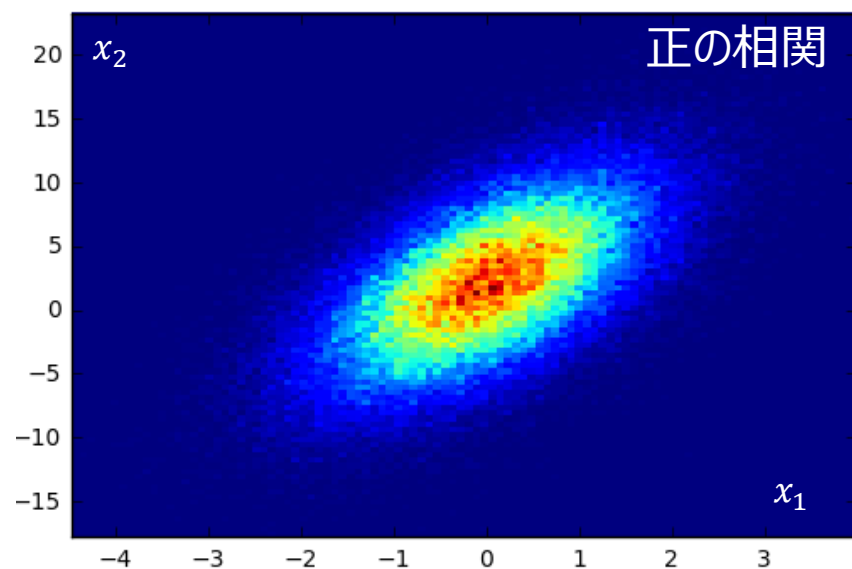
	50kg以下	50-60kg	60-70kg	70kg以上
150cm以下	2	1	0	0
150-160cm	4	12	5	1
160-170cm	1	10	13	2
170cm以上	0	2	4	4

● 棒グラフタイプ



先述の通り, 「死角」があり, さらに高さを直接比較しにくいのでお勧めしない

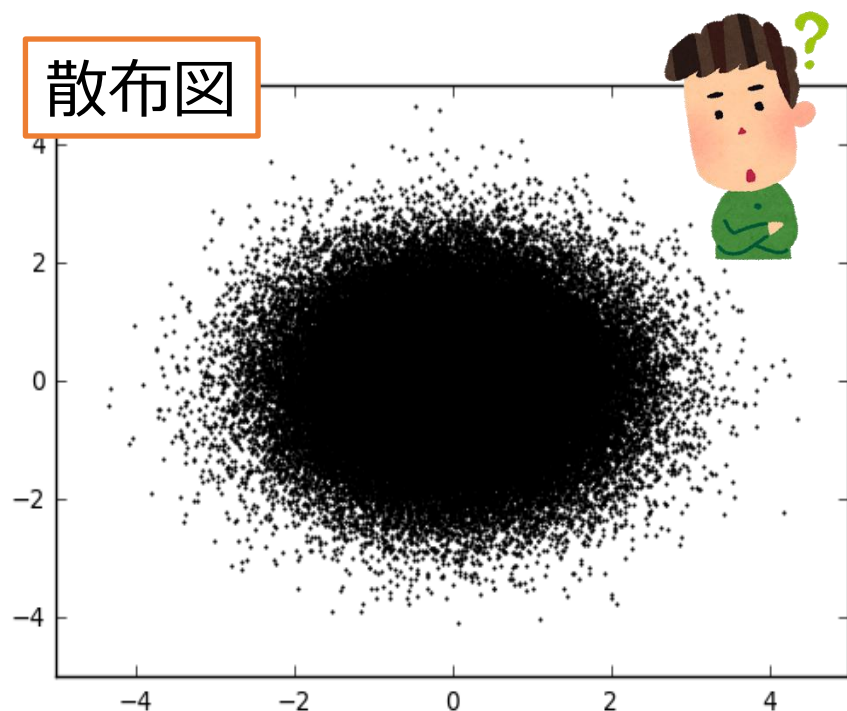
相関がある2次元データの2次元ヒストグラム



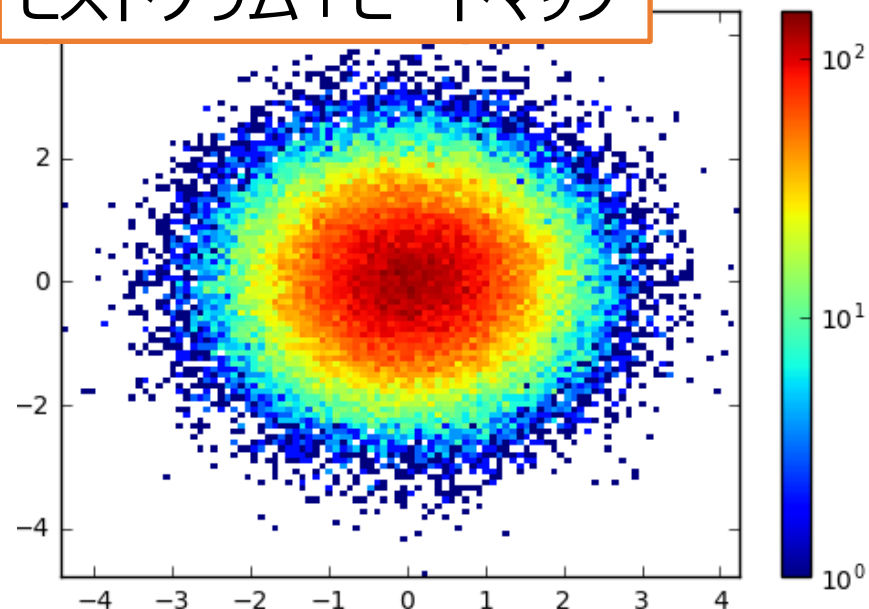
散布図 vs. ヒートマップ

- データが多いとヒートマップ化が必要

散布図



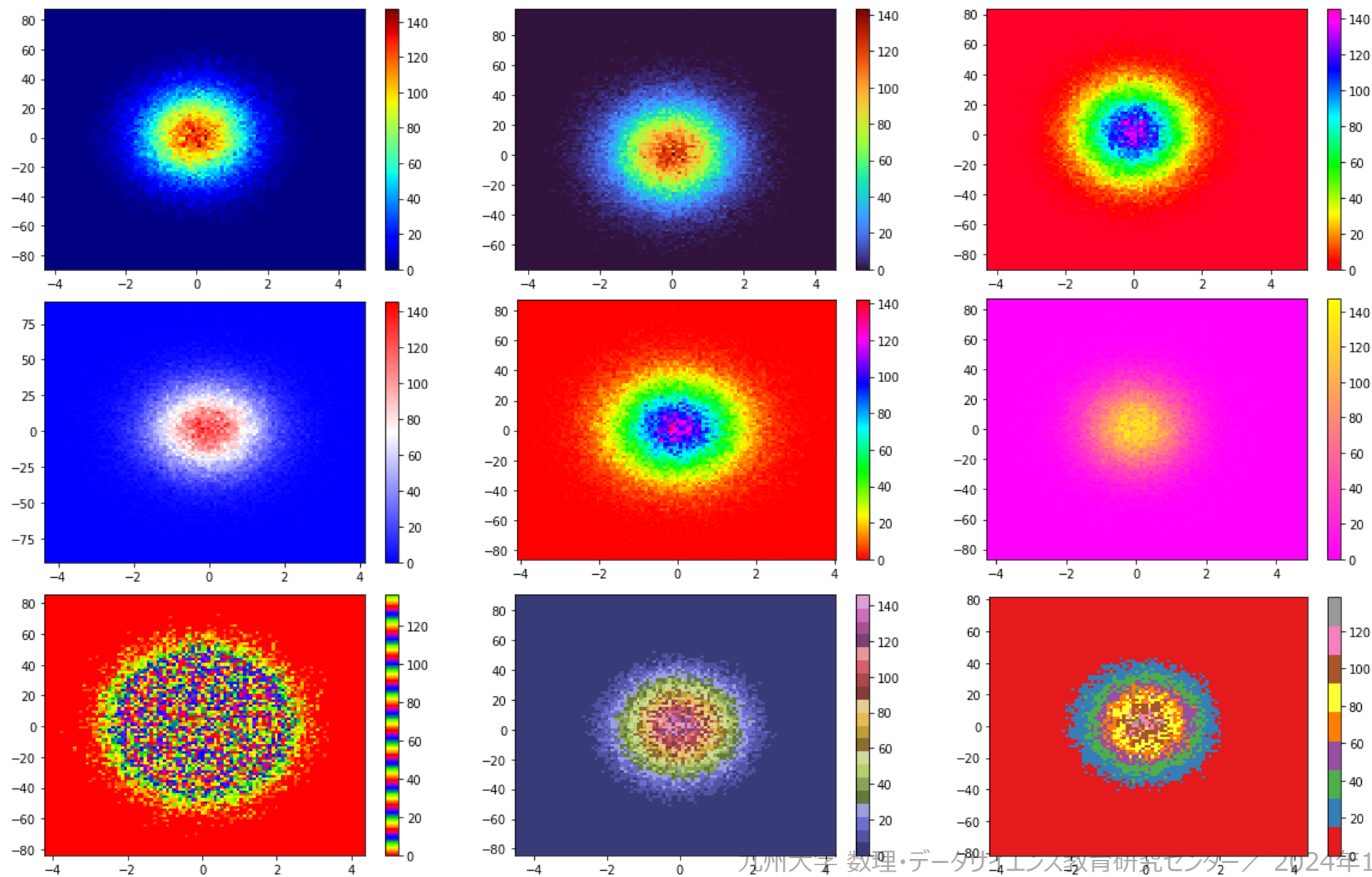
ヒストグラム+ヒートマップ



こんなヒートマップの使い方



カラーマップ（カラースケール）の選択に注意： 同じデータでも見え方が全然変わってしまう

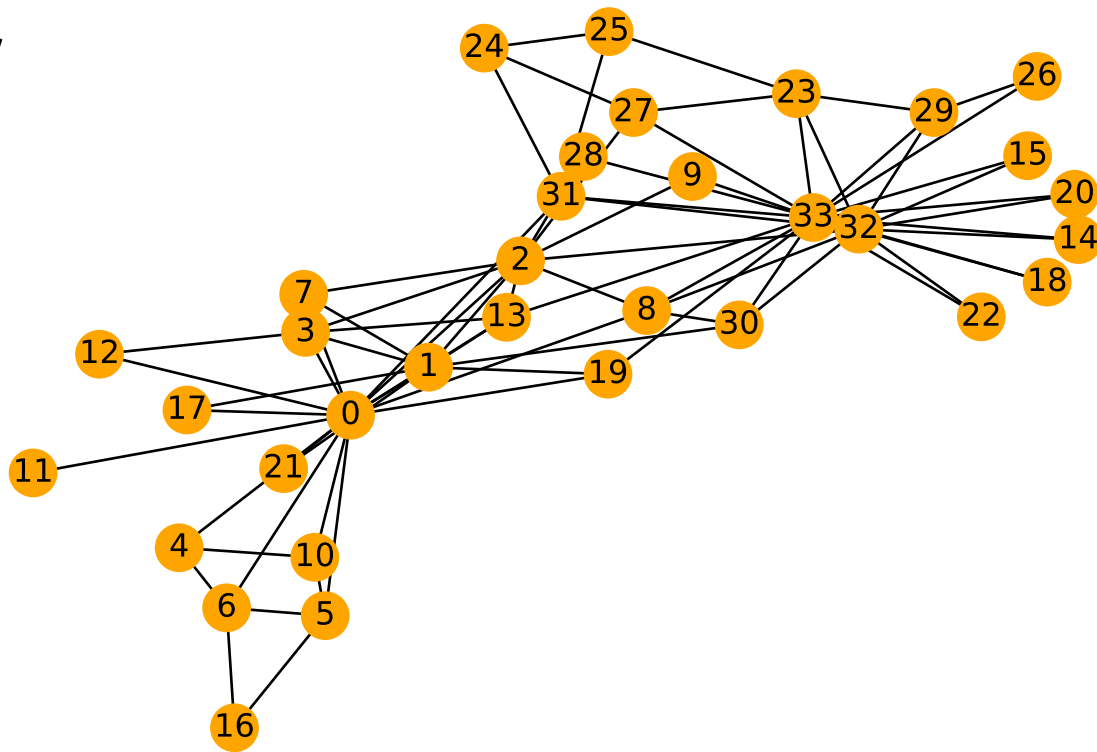


その他の可視化

関係性の可視化

- ネットワークによる人間関係の可視化の例

- ○印 = 空手クラブのメンバー
- 線 = 仲良し

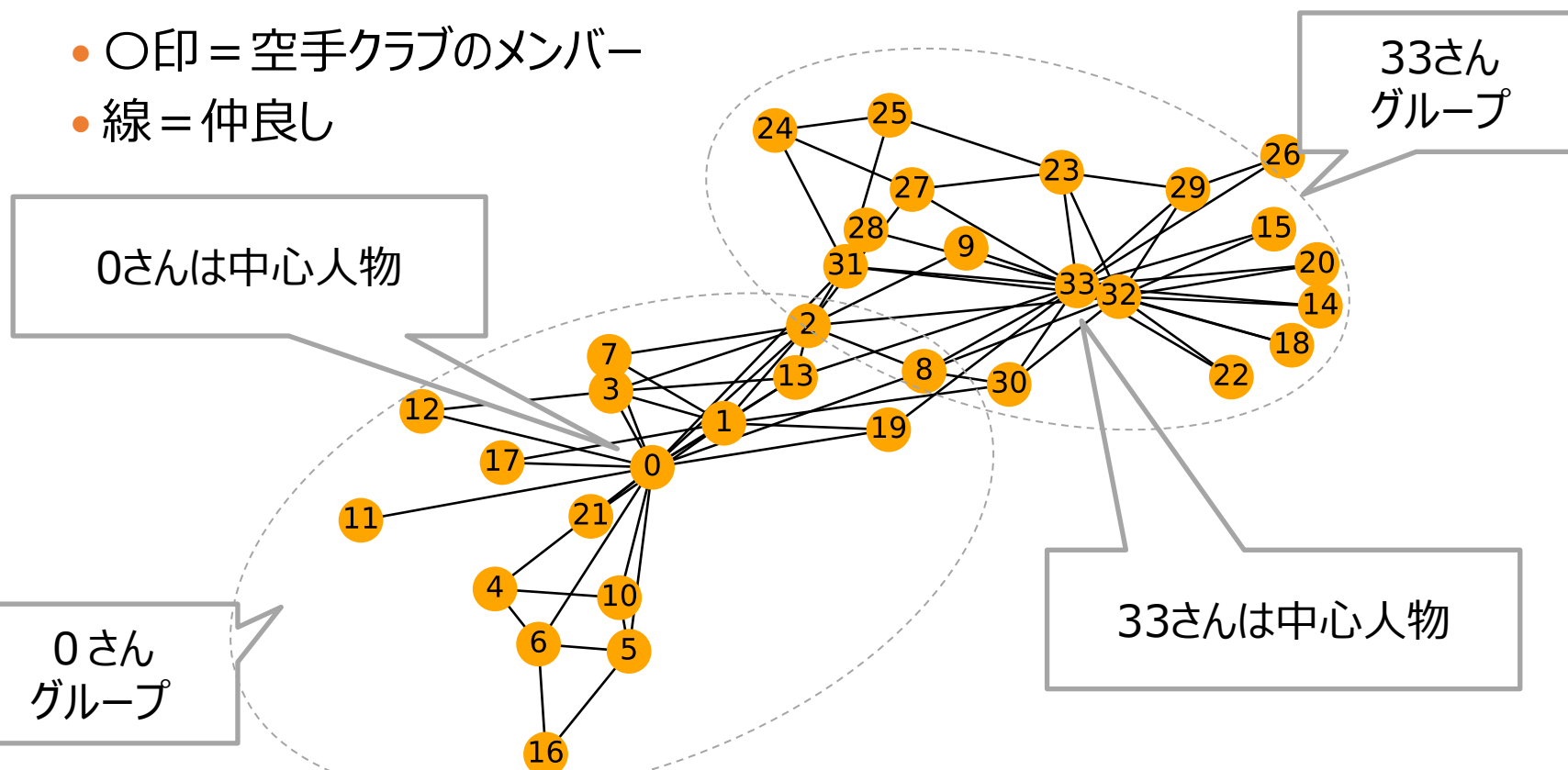


- このように、およそ 2 グループに分かれる様子がわかる

関係性の可視化

- ネットワークによる人間関係の可視化の例

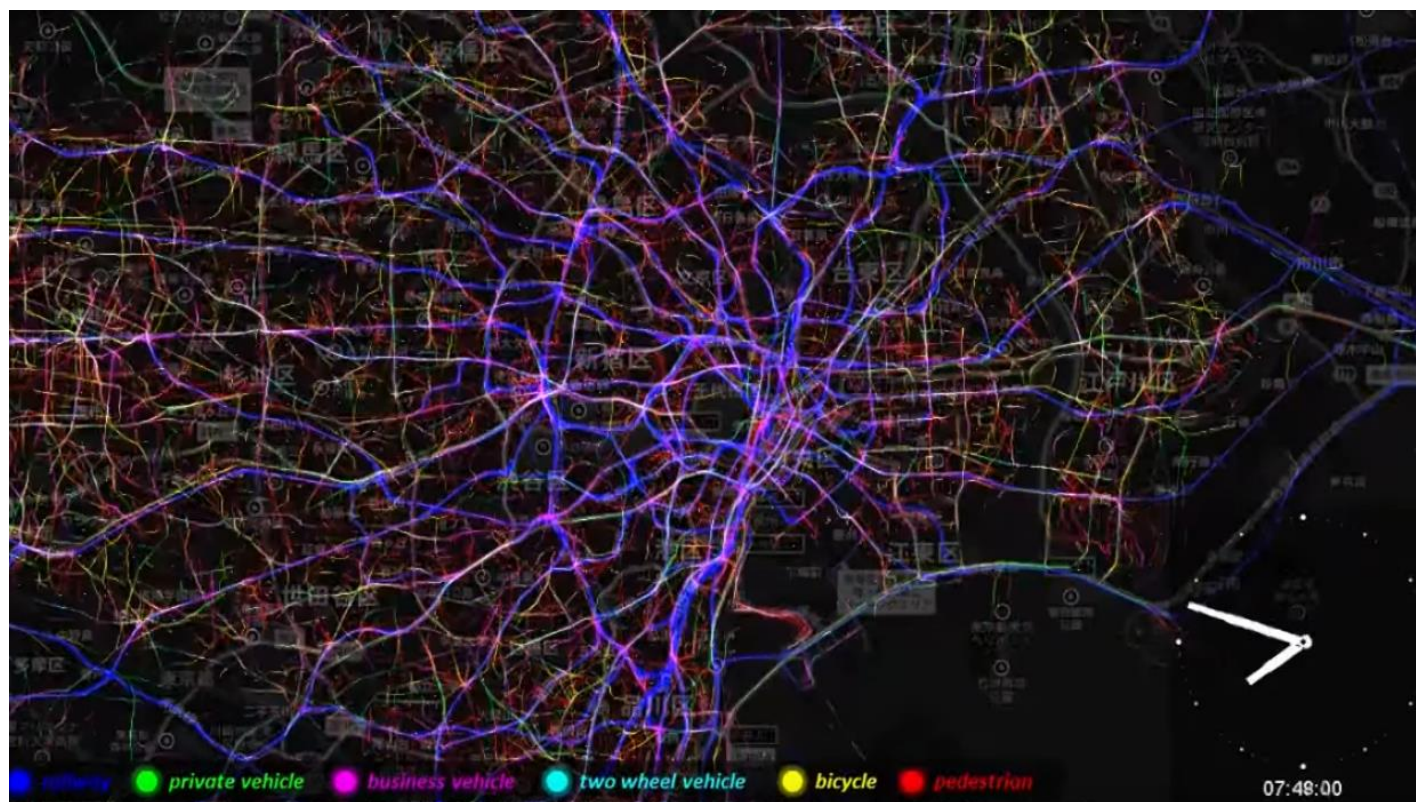
- 印 = 空手クラブのメンバー
- 線 = 仲良し



- このように、およそ2グループに分かれる様子がわかる

挙動・軌跡の可視化

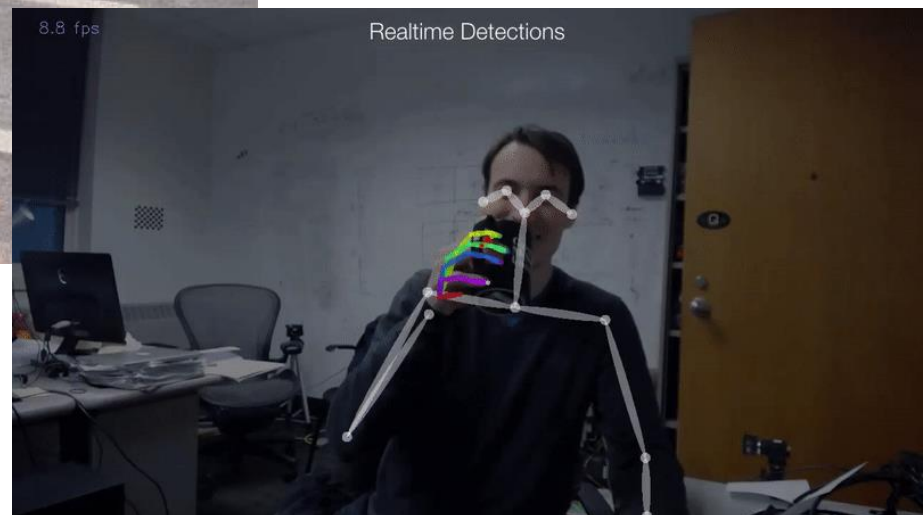
- 人々の移動の様子



- 是非動画で→ <https://pflow.csis.u-tokyo.ac.jp/data-visualization/visualization/>

リアルタイム可視化： 変化の様子を時々刻々と表示する

- Openposeを使った人々の動きのリアルタイム可視化

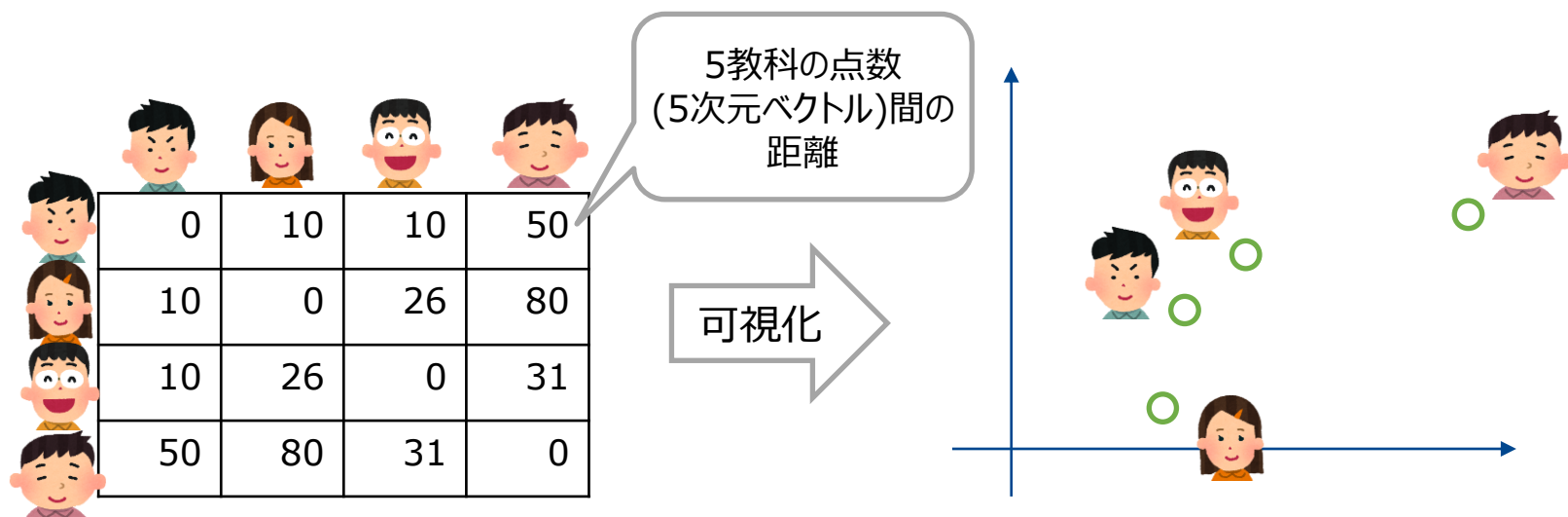


【付録 1】 多次元尺度構成法とt-SNE

データ間の近さを「できるかぎり」保ったまま， 2次元平面に押しつぶす

多次元尺度構成法 (Multi-dimensional Scaling)

- データ間の距離が与えられれば, なるべくそれらの距離を保つように, 2次元(or 3次元) にプロットする



- これ ↓ が小さくなるよう可視化面での位置 $\{x_i\}$ を求める (バリエーションあり)

$$\left(\sum_{i,j} (d_{i,j} - \|x_i - x_j\|)^2 \right)^{1/2}$$

元の距離 (上の表)

可視化に用いる低次元空間での距離

多次元尺度構成法のイメージ図

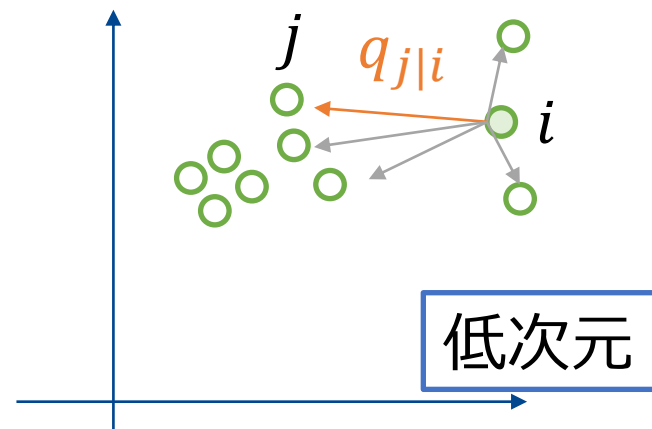
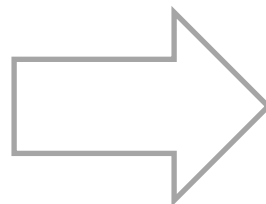
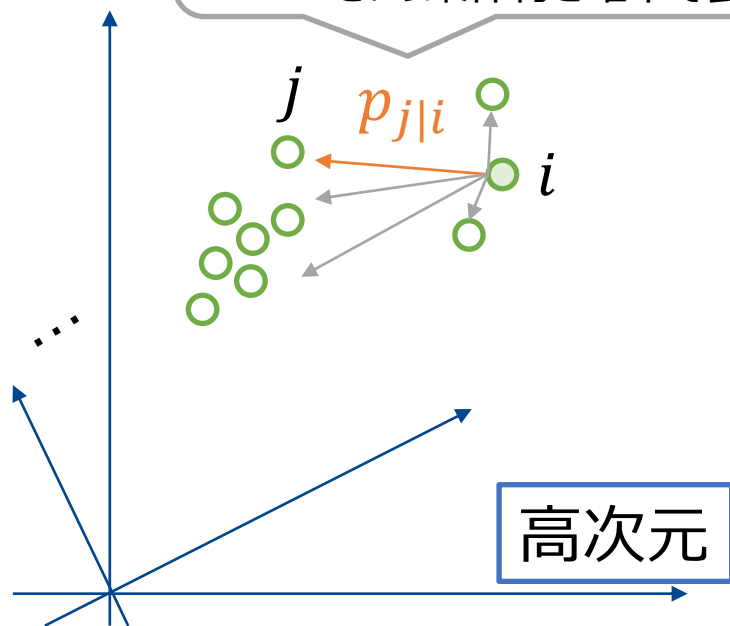
- 4つの3次元データを2次元にする場合



t-SNE [van der Maaten+, JMLR, 2008]

t-distributed Stochastic Neighbor Embedding, t 分布型確率的近傍埋め込み

i からみた j の類似度
「 i が存在したときに j はどれくらい出やすいか」
という条件付き確率で表す



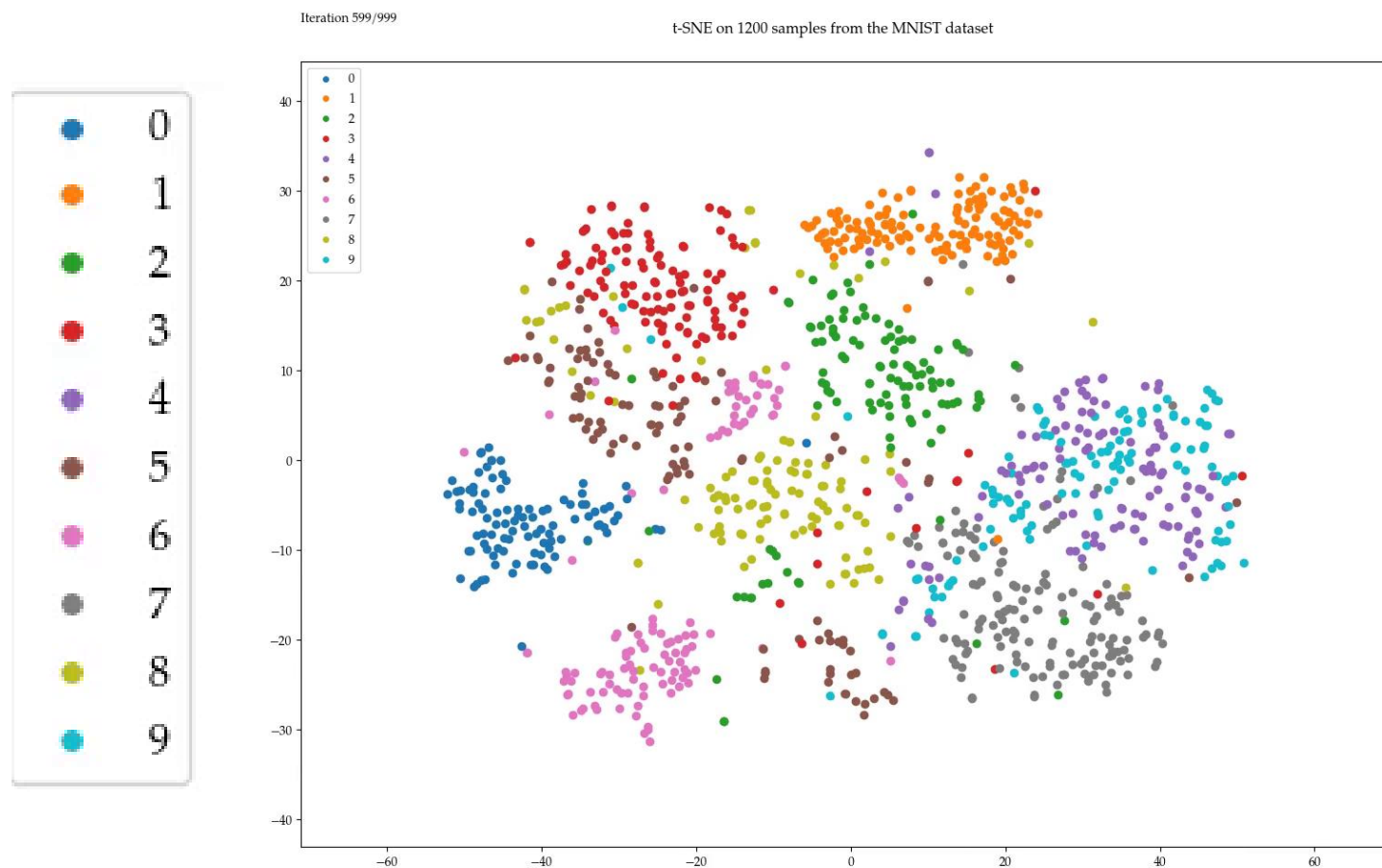
すべての i, j について $p_{j|i} \sim q_{j|i}$ となるように,
低次元空間上での i, j の位置を(徐々に)最適化

(式はずいぶん違うが, 根本となっている考え方は多次元尺度構成法と似ている)

t-SNE [van der Maaten+, JMLR, 2008]

t-distributed Stochastic Neighbor Embedding, t 分布型確率的近傍埋め込み

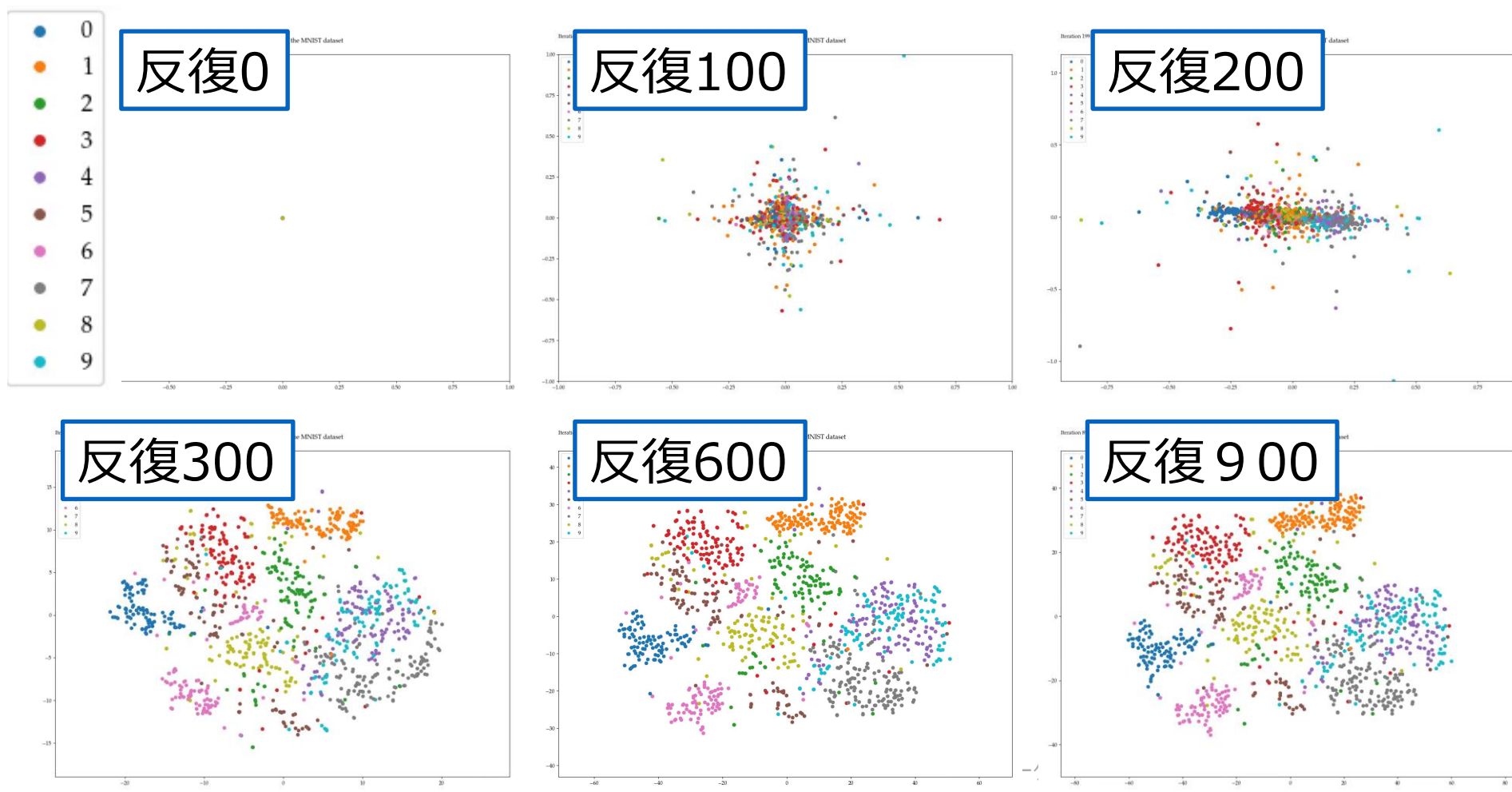
- 手書き数字画像データ(MNIST)をt-SNE可視化した場合の動画



t-SNE [van der Maaten+, JMLR, 2008]

t-distributed Stochastic Neighbor Embedding, t 分布型確率的近傍埋め込み

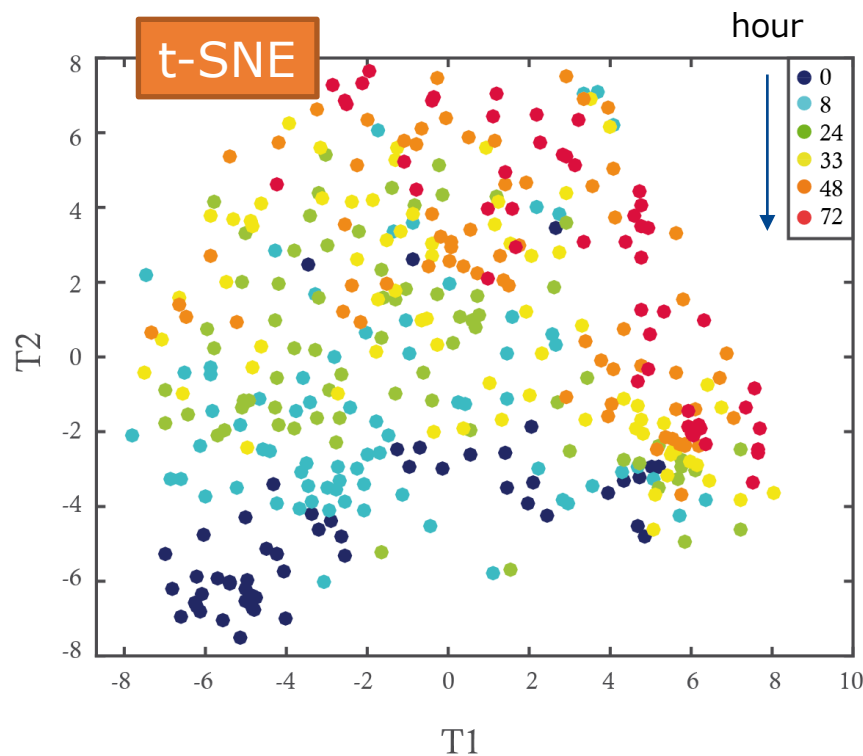
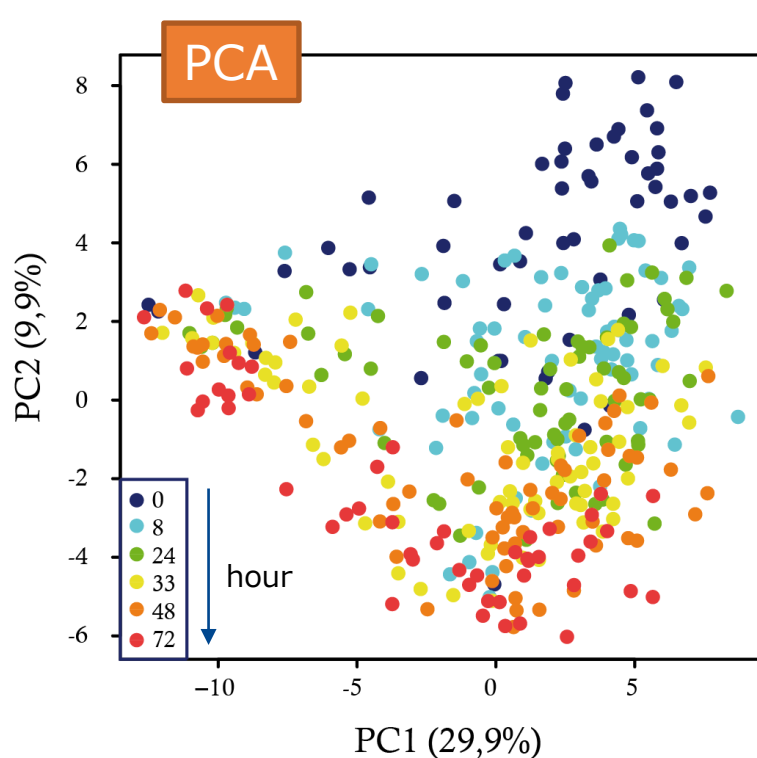
- 手書き数字画像データ(MNIST)をt-SNE可視化した場合の静止画



PCA(まっすぐ射影) vs t-SNE(ひん曲げる)

● 90次元の遺伝子発現の時間的推移

[Richard+, PLOS Biology, 2016] <https://doi.org/10.1371/journal.pbio.1002585>



● 同じデータなのに，可視化法によって違った分布に見える点に注意