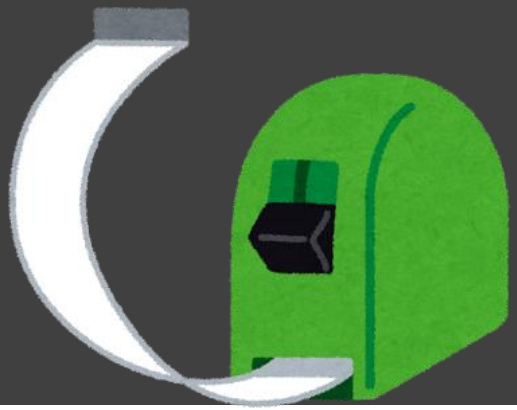
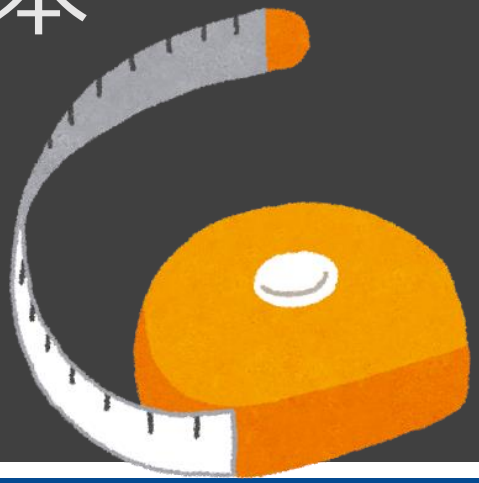
# データサイエンス概論I＆II

## データ間の距離と類似度

九州大学 数理・データサイエンス教育研究センター

# 「データ間の距離」の基本

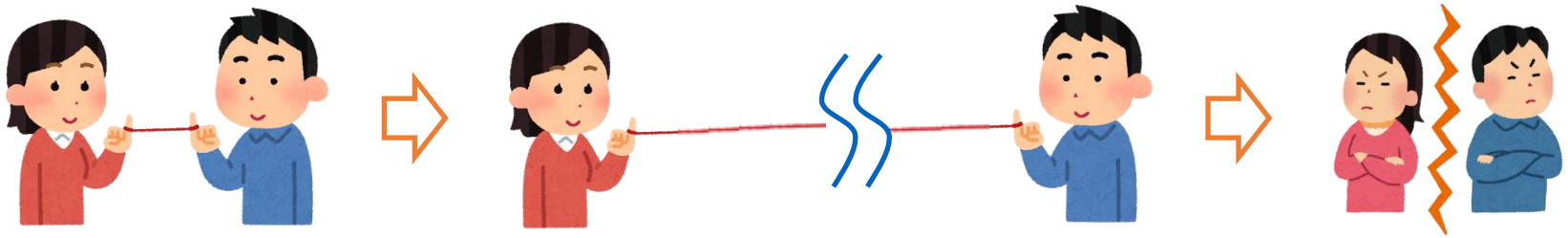




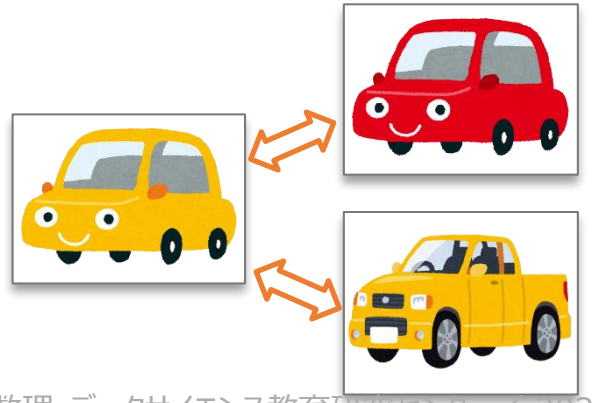
「A君って，タレントの○○に似てるよね？」
「えー，全然似てないよ．むしろ△△でしょう．」
「いや，自分では□□似だと思ってるんだけど．」

# データ解析における「距離」とは？

- 日常会話における「距離」
  - A地点とB地点がどれぐらい離れているか？ （単位：mとかkmとか）
  - Aさんの気持ちとBさんの気持ちがどれぐらい離れているか？

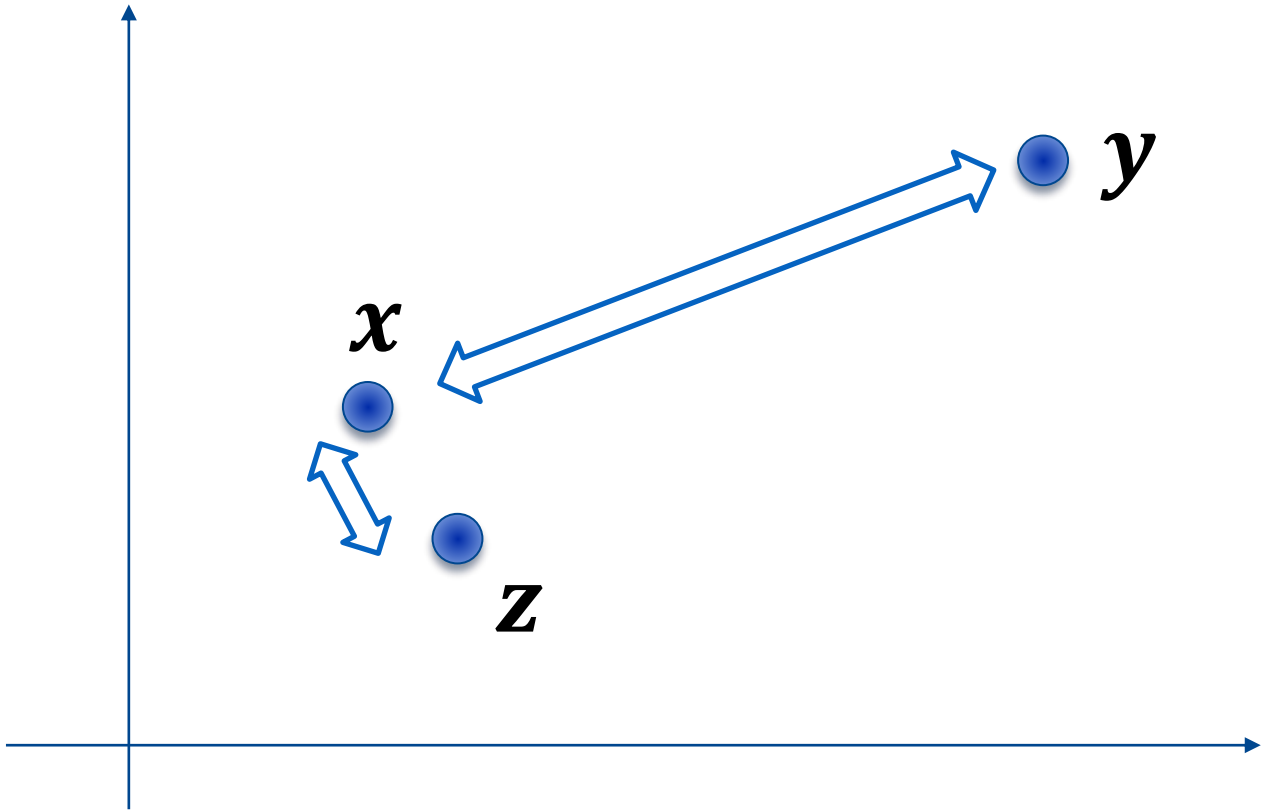


- データ解析における「距離」はもっと一般的
  - 要するにデータ間の差異 （似てない具合）
  - 距離が小さい2データは「似ている」
  - 単位がある場合もない場合も



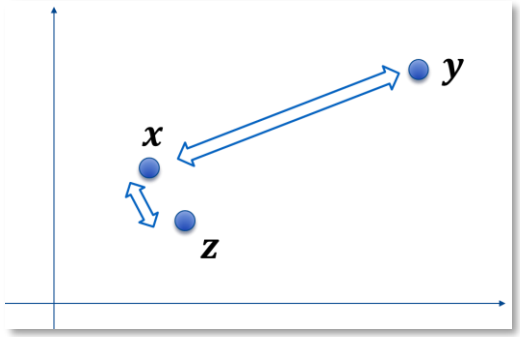
# 普通に考える「データ間の距離」：2データがどれぐらい違うか？(＝離れているか？)

- $\boldsymbol{x}$にとって，　$\boldsymbol{y}$は結構違っていて，　$\boldsymbol{z}$は似ている

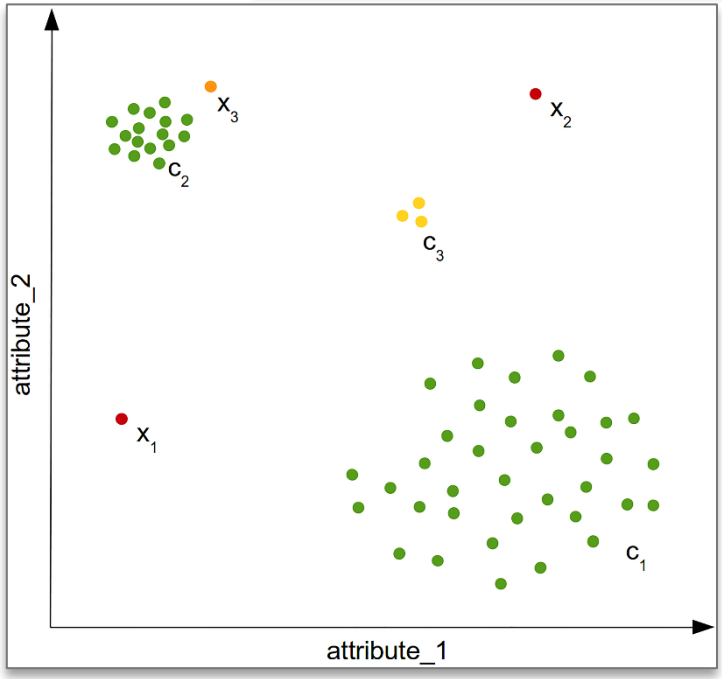


# 距離がわかると何に使えるか？ 実は超便利！（1/2）

- データ間の比較が定量的にできる
  - 「$\boldsymbol{x}$と$\boldsymbol{y}$は全然違う/結構似ている」「$\boldsymbol{x}$と$\boldsymbol{y}$は28ぐらい違う」
  - 「$\boldsymbol{x}$にとっては，$\boldsymbol{y}$よりも$\boldsymbol{z}$のほうが似ている」




- データ集合のグルーピングができる
  - 「近く」のデータどうしでグループを作る
  - 「クラスタリング」と呼ばれる




[Goldstein, Uchida, PLoSONE, 2016]

- データの異常度が測れる
  - 「近く」にデータがたくさんあれば正常，一つもなければ異常

# 距離がわかると何に使えるか？<br>実は超便利！ (2/2)

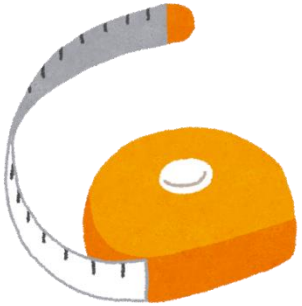
- データの「認識」ができる
  - 登録されている画像データ中で，画像$\boldsymbol{x}$に最も似ているものは「リンゴ」だった<br>→ 「画像 $\boldsymbol{x}$ はリンゴ」と判断

- 近似精度が測れる
  - データ $\boldsymbol{x}$を $\boldsymbol{y}$で近似(代用)した時の誤差 → $\boldsymbol{x}$ と $\boldsymbol{y}$の距離に等しい

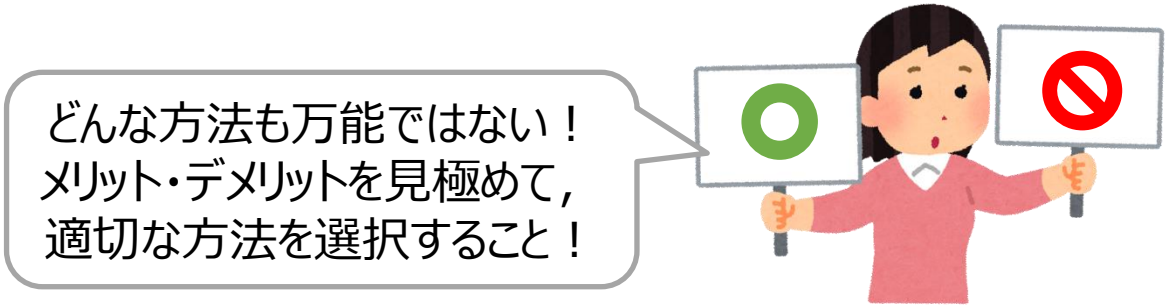
- … and more!

# 「距離」の話を通して学んで頂きたいこと

- 距離は「データ解析の基本」である！
- **距離は１種類ではない！**
- 距離が変われば，データ解析結果は「まるっきり」変わる



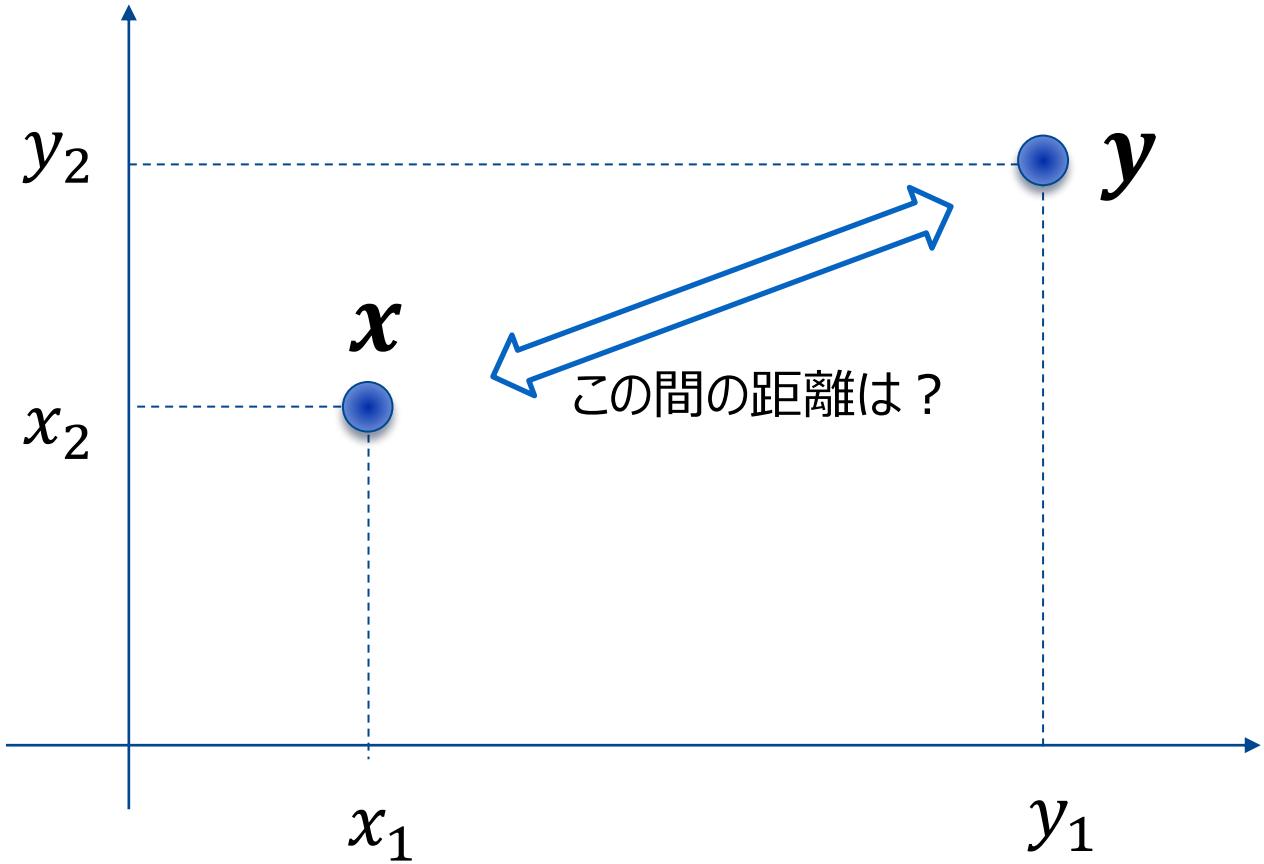
- 解析問題の性質に合致した「距離」を選ぶ必要がある
  - 様々な距離の原理，メリット・デメリットも理解しておこう



人の名前(後述)

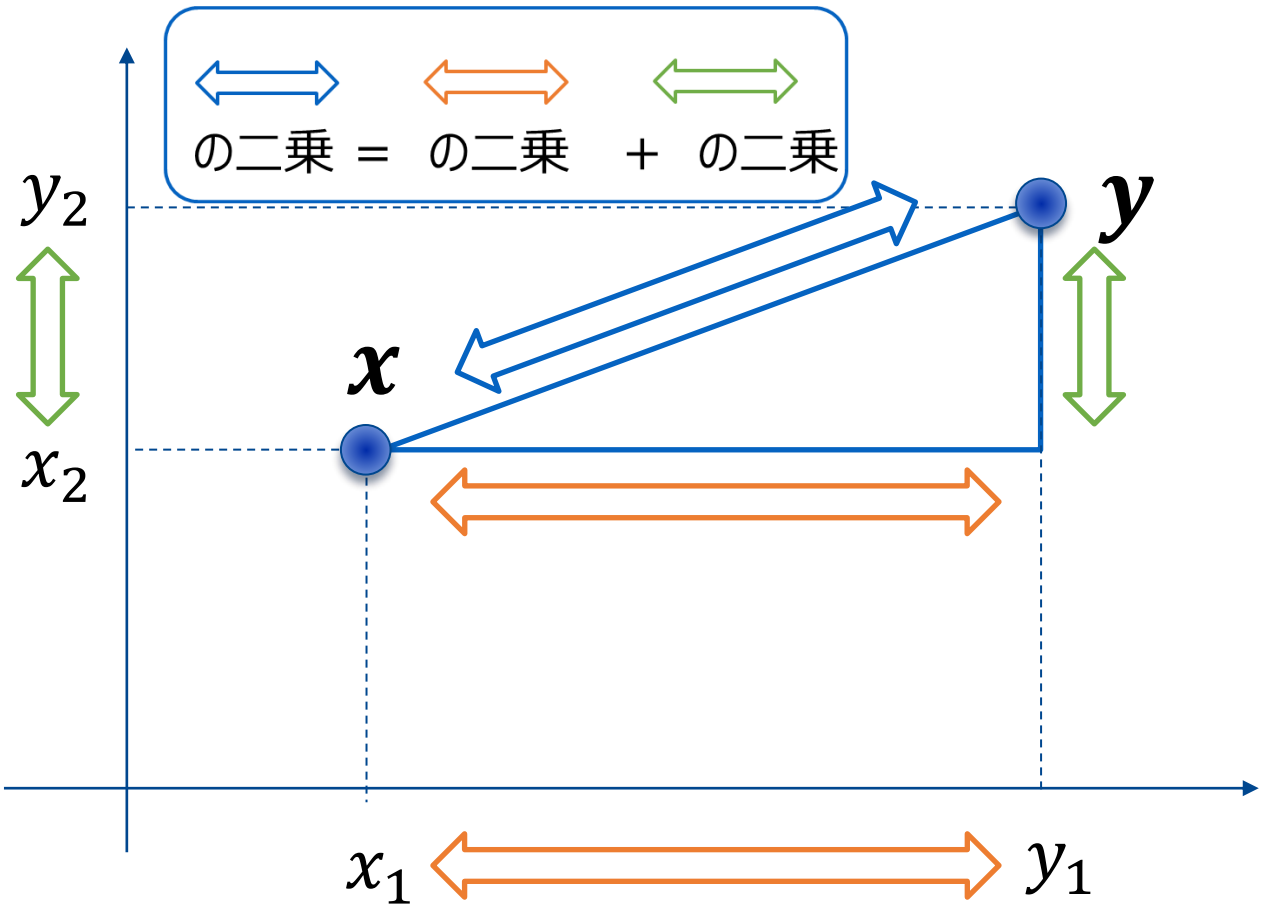
# 最も代表的な距離：ユークリッド距離（1）

- 地図上の2点　$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$



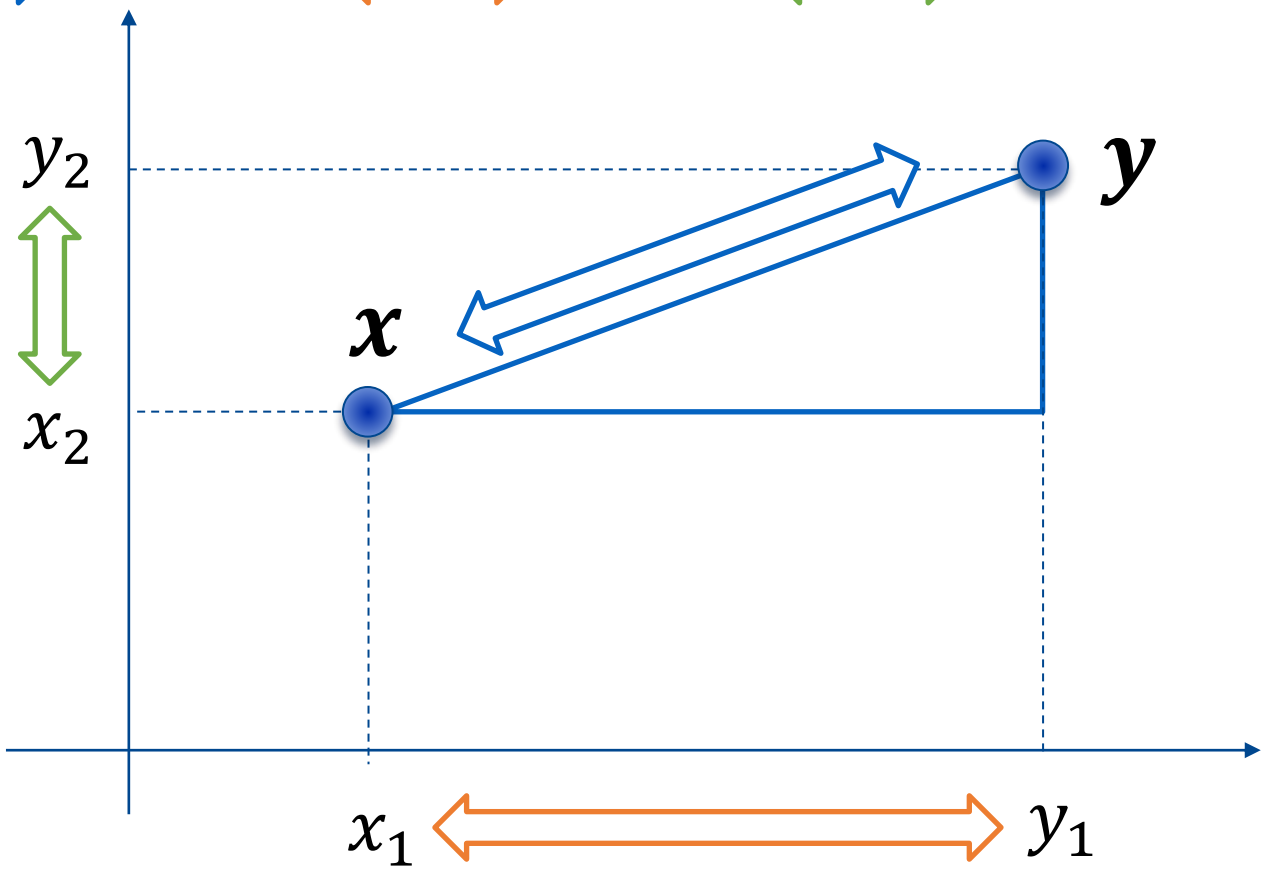
# 最も代表的な距離：ユークリッド距離（2）

- ご存じ「三平方の定理」（ピタゴラスの定理）



# 最も代表的な距離：ユークリッド距離（3）

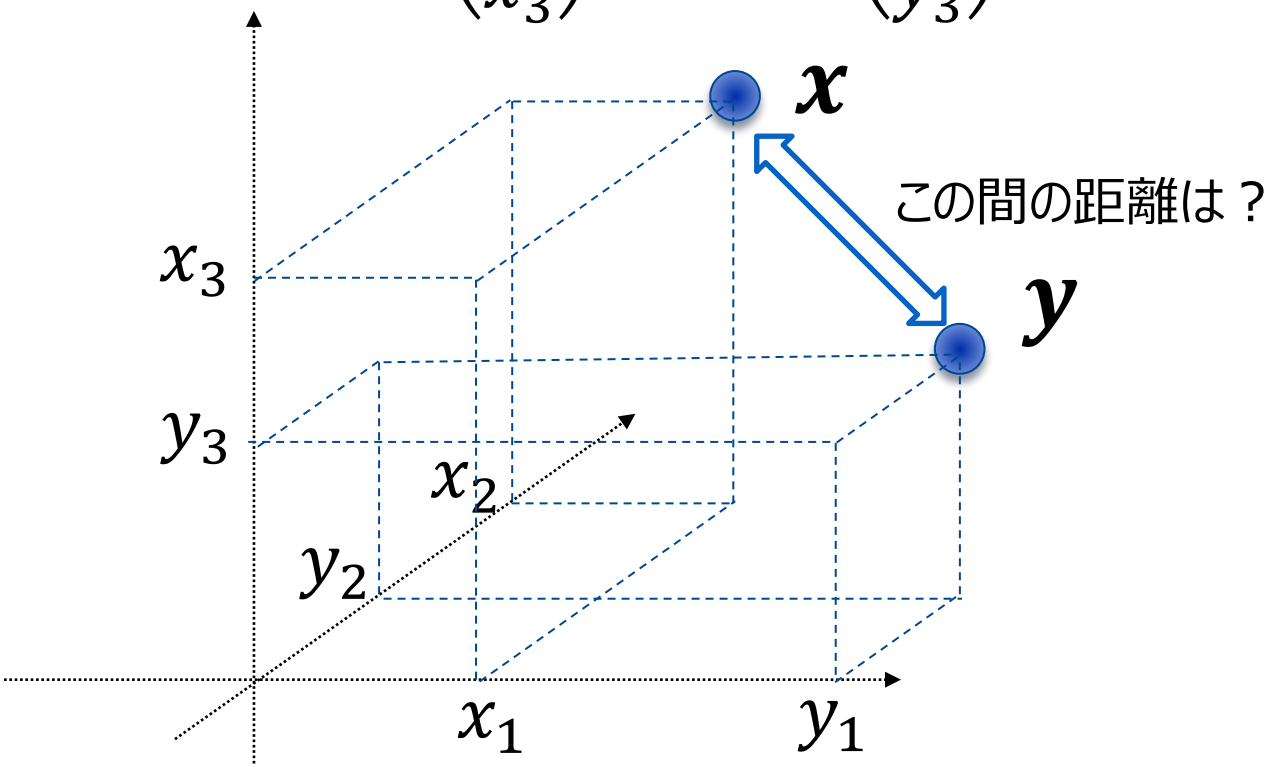
- $\boldsymbol{x}$と$\boldsymbol{y}$の距離の二乗 $=(x_1-y_1)^2+(x_2-y_2)^2$



# 最も代表的な距離：ユークリッド距離（4）

- 3 次元だとどうなる？

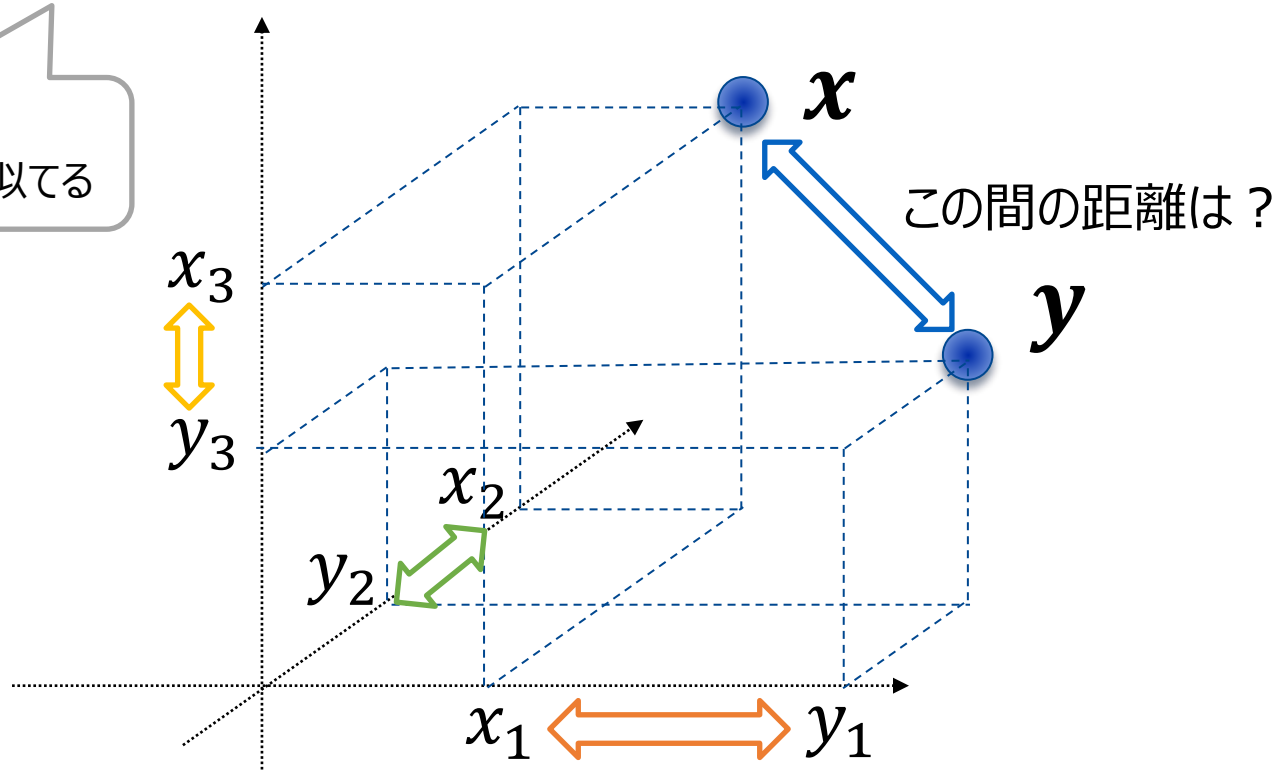
$$\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$$



# 最も代表的な距離：ユークリッド距離（5）

- $\boldsymbol{x}$と$\boldsymbol{y}$の距離の二乗 $=(x_1-y_1)^2+(x_2-y_2)^2+(x_3-y_3)^2$

なんかやっぱり
ピタゴラスの定理に似てる



# 最も代表的な距離：ユークリッド距離 (6)

- 2次元の場合

$$\boldsymbol{x}\text{と}\boldsymbol{y}\text{の距離の二乗} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} x_1 - y_1 \\ x_2 - y_2 \end{pmatrix}$$

$\boldsymbol{x}$ $\boldsymbol{y}$ $\boldsymbol{x} - \boldsymbol{y}$

二乗 ＋ 二乗

- 3次元の場合

$$\boldsymbol{x}\text{と}\boldsymbol{y}\text{の距離の二乗} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} - \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_1 - y_1 \\ x_2 - y_2 \\ x_3 - y_3 \end{pmatrix}$$

$\boldsymbol{x}$ $\boldsymbol{y}$ $\boldsymbol{x} - \boldsymbol{y}$

二乗 ＋ 二乗 ＋ 二乗

# 最も代表的な距離：ユークリッド距離（7）

- $d$次元の場合

$\boldsymbol{x}$と$\boldsymbol{y}$の距離の二乗

$$= \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} - \begin{pmatrix} y_1 \\ \vdots \\ y_d \end{pmatrix} = \begin{pmatrix} x_1 - y_1 \\ \vdots \\ x_d - y_d \end{pmatrix}$$

$\boldsymbol{x}$　$\boldsymbol{y}$　$\boldsymbol{x} - \boldsymbol{y}$

二乗 ＋ ⋮ ＋ 二乗

というわけで，何次元ベクトルでも距離は計算可能

もちろん1次元ベクトル(数値)間の距離も計算可能

# 最も代表的な距離：ユークリッド距離 (8)

- 簡略表現法

$$\boldsymbol{x}\text{と}\boldsymbol{y}\text{の距離の二乗} = (\boldsymbol{x} - \boldsymbol{y})^2$$

$$\begin{pmatrix} x_1 - y_1 \\ \vdots \\ x_d - y_d \end{pmatrix} \quad \begin{matrix} \text{二乗} \\ + \\ \vdots \\ + \\ \text{二乗} \end{matrix}$$

# 最も代表的な距離：ユークリッド距離 (9)

- これでようやくユークリッド距離

$$\boldsymbol{x}\text{と}\boldsymbol{y}\text{の距離の二乗} = (\boldsymbol{x}-\boldsymbol{y})^2$$

$$\boldsymbol{x}\text{と}\boldsymbol{y}\text{の距離} = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^2} = \|\boldsymbol{x}-\boldsymbol{y}\|$$

なんだこの記号？
(次スライド)

$d$次元ベクトル $\boldsymbol{x}$と$\boldsymbol{y}$ の間のユークリッド距離

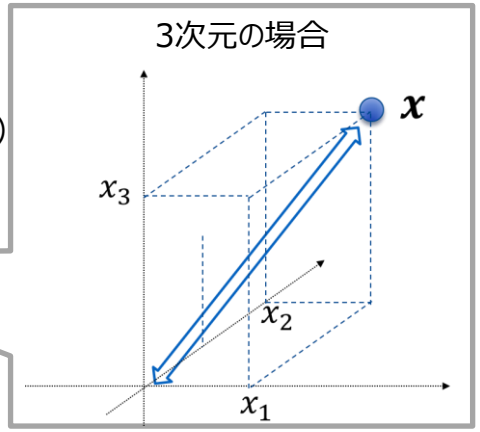
# 参考：なんだこの二重絶対値||·||は？

- $\|\boldsymbol{x}\|$はベクトル$\boldsymbol{x}$の長さを表すんです
  - ベクトル$\boldsymbol{x}$の「ノルム」とも言います！

- ベクトル$\boldsymbol{x}$の長さは(実はノルムにもいろいろあるんですが，そんなことまずは気にせずに考えれば)

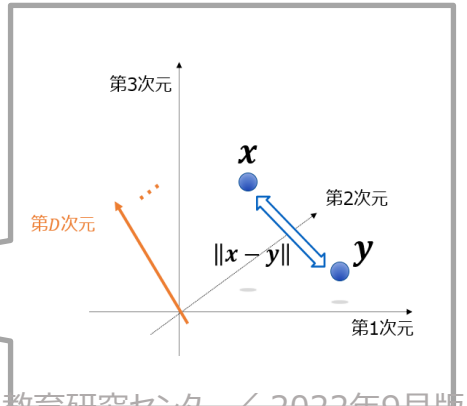
$$\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_d^2}$$

となります

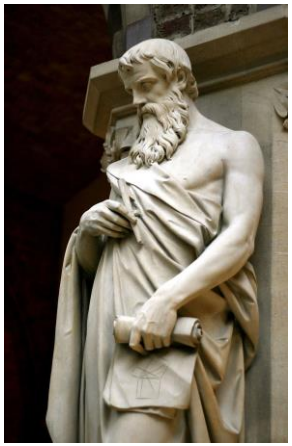



- だから$\|\boldsymbol{x}-\boldsymbol{y}\|$は$\boldsymbol{x}$と$\boldsymbol{y}$の差の長さ，
すなわち距離ってわけです

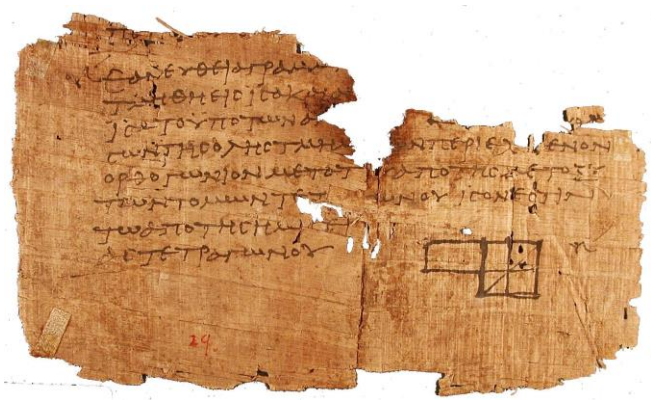
$$\|\boldsymbol{x}-\boldsymbol{y}\| = \sqrt{(\boldsymbol{x}-\boldsymbol{y})^2}$$



# 参考：ユークリッド＝幾何学の父



@エジプト
BC330～275年頃？

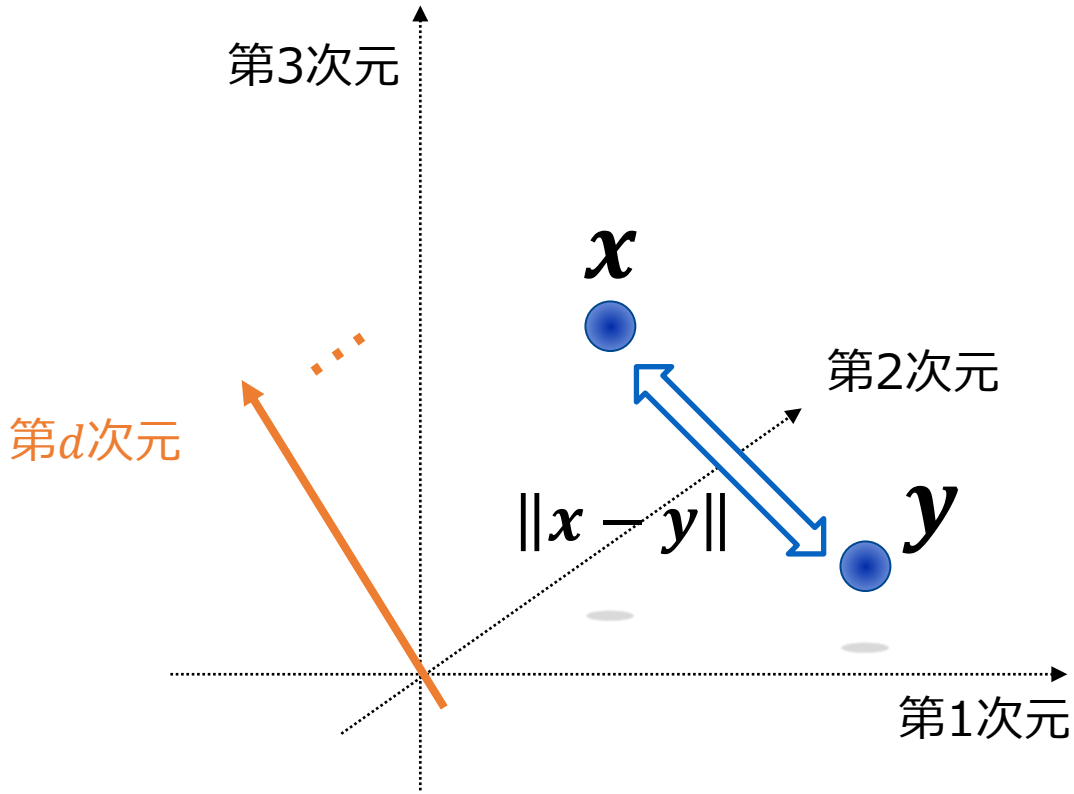


ユークリッド原論

- ユークリッド原論にある5つの公準(≒公理)
  - 第1公準 ： 点と点を直線で結ぶ事ができる
  - 第2公準 ： 線分は両側に延長して直線にできる
  - 第3公準 ： 1点を中心にして任意の半径の円を描く事ができる
  - 第4公準 ： 全ての直角は等しい（角度である）
  - 第5公準 ： 1つの直線が2つの直線に交わり、同じ側の内角の和が2つの直角より小さいならば、この2つの直線は限りなく延長されると、2つの直角より小さい角のある側において交わる（≒平行線でない2直線は1点で交わる）

# 最も代表的な距離：ユークリッド距離（10）

- 図示するとやっぱりこんな感じ



# 練習：2つのデータ間のユークリッド距離を求めよう

$\boldsymbol{x} = (3), \boldsymbol{y} = (6)$のとき $\|\boldsymbol{x} - \boldsymbol{y}\|$は？

$\boldsymbol{x} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} 6 \\ 1 \end{pmatrix}$のとき $\|\boldsymbol{x} - \boldsymbol{y}\|$は？

$\boldsymbol{x} = \begin{pmatrix} 3 \\ 5 \\ 2 \end{pmatrix}, \boldsymbol{y} = \begin{pmatrix} 6 \\ 1 \\ 2 \end{pmatrix}$のとき $\|\boldsymbol{x} - \boldsymbol{y}\|$は？

# これで画像間の距離(似てる具合)も測れます

どちらも1000x1000画素の画像

100万次元ベクトル $\boldsymbol{x}$

100万次元ベクトル $\boldsymbol{y}$

画像間距離

$$\|\boldsymbol{x}-\boldsymbol{y}\|$$

おっ、これで画像認識AIができそう

尺度の異なる数値が組み合わされたデータに関する距離の計算は要注意

実データ間の距離を測る際の留意点(1/3)

# (身長, 体重)データ間のユークリッド距離

$\begin{pmatrix} 体重(kg) \\ 身長(cm) \end{pmatrix}$とする．このとき

$\boldsymbol{x} = \begin{pmatrix} 60 \\ 150 \end{pmatrix}$と似ているのは，

$\boldsymbol{y} = \begin{pmatrix} 30 \\ 150 \end{pmatrix}$と $\boldsymbol{z} = \begin{pmatrix} 60 \\ 153.1 \end{pmatrix}$のどっち？

それはやはり$\boldsymbol{z}$さんのほうが似てる(30kg差は大きい)

# (身長, 体重)データ間のユークリッド距離

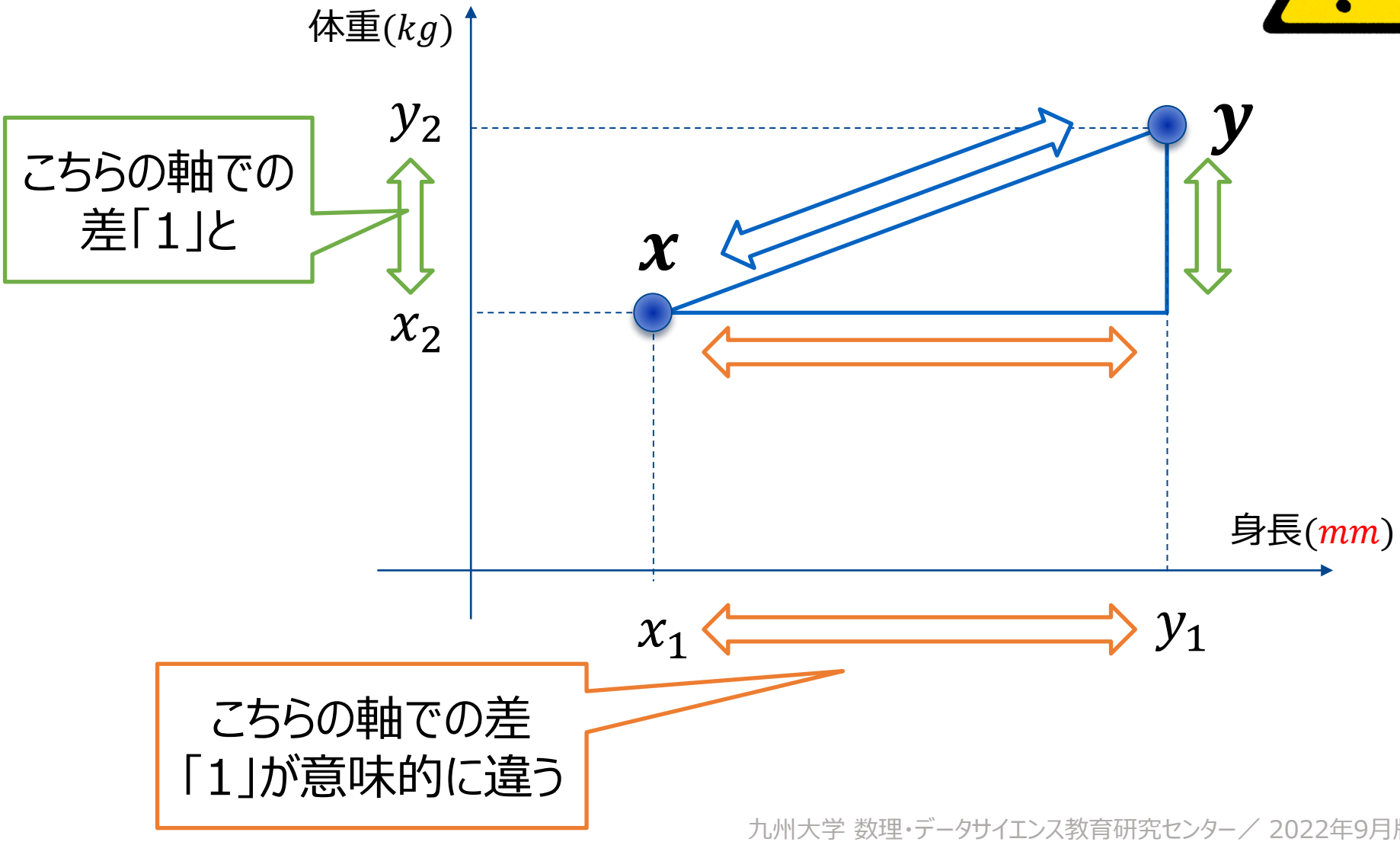
$\begin{pmatrix} \text{体重}(kg) \\ \text{身長}(mm) \end{pmatrix}$とする．このとき

$\boldsymbol{x} = \begin{pmatrix} 60 \\ 1500 \end{pmatrix}$と似ているのは,

$\boldsymbol{y} = \begin{pmatrix} 30 \\ 1500 \end{pmatrix}$と $\boldsymbol{z} = \begin{pmatrix} 60 \\ 1531 \end{pmatrix}$のどっち？

単位が変わるだけで思ってもないことに..

# というわけで，尺度の異なる数値が組み合わされたデータに関する距離の計算は要注意！



# さっきの画像間距離，なぜ大丈夫なのか？

100万次元ベクトル $\boldsymbol{x}$

100万次元ベクトル $\boldsymbol{y}$

100万個の
どの要素も全部
「画素値」

画像間距離
$\|\boldsymbol{x}-\boldsymbol{y}\|$

100万個の
どの要素も全部
「画素値」

安心してそのまま距離計算可能

# この辺，結構悩ましい:
# (数学点数, 古文点数)データ間のユークリッド距離

$$\begin{pmatrix} \text{数学の点数(100点満点)} \\ \text{古文の点数(50点満点)} \end{pmatrix}$$

について普通に距離計算していいのか？

解釈1

同じどちらも「点数」なので，そのまま計算してOK

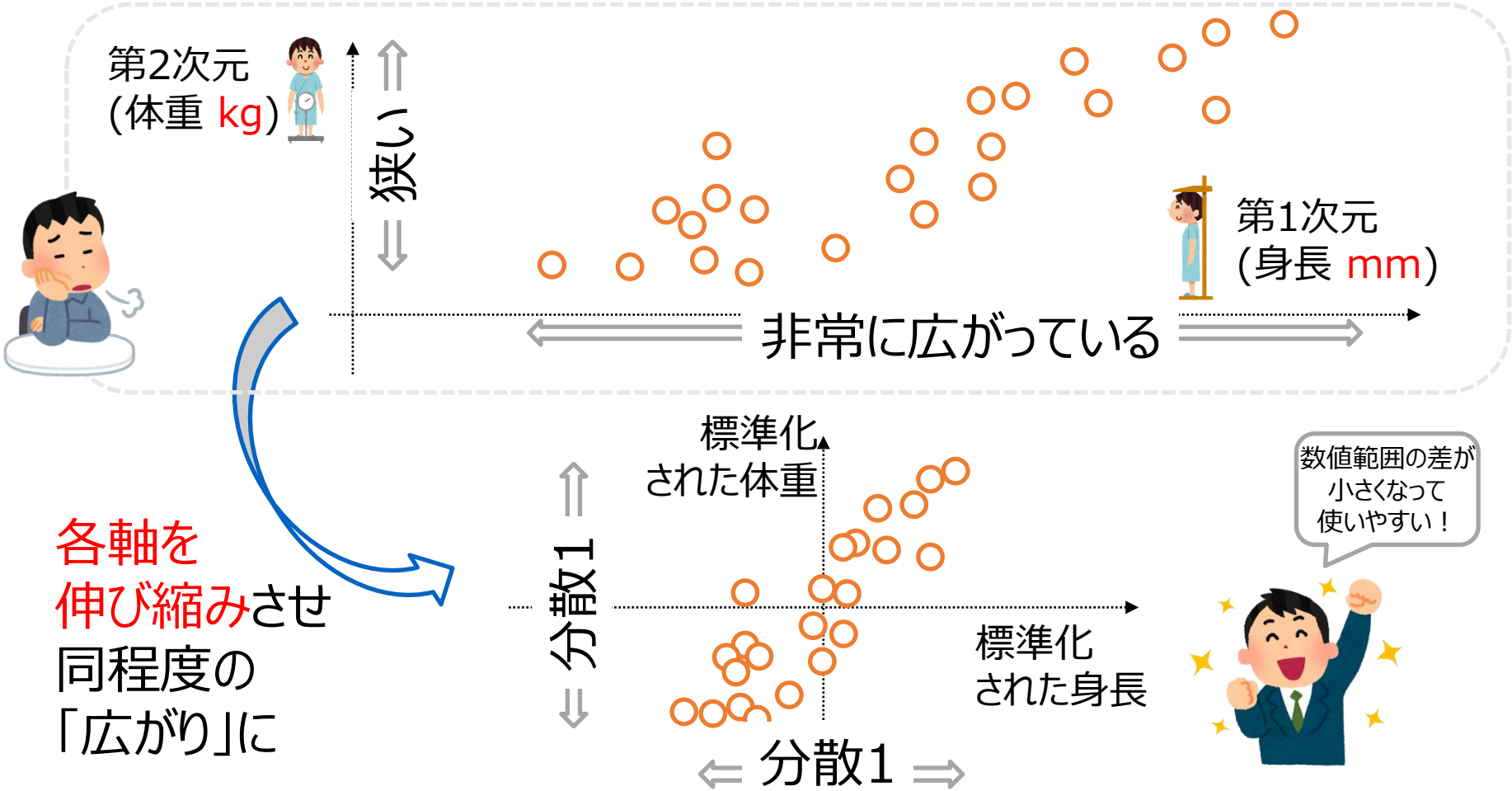
解釈2

古文を２倍にしてどちらも100点満点にしてから距離計算すべき

どちらかが「正しい」という証明はできない
(ケースバイケースで使い分けするしかない)

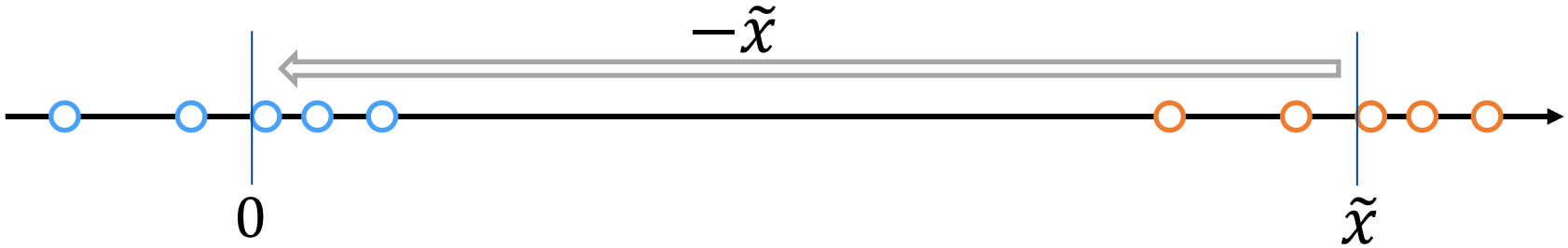
# 物理的に意味の異なる数値を扱う場合に関して リーズナブルな方法：標準化

- 平均をゼロ，分散を1に「標準化」する！

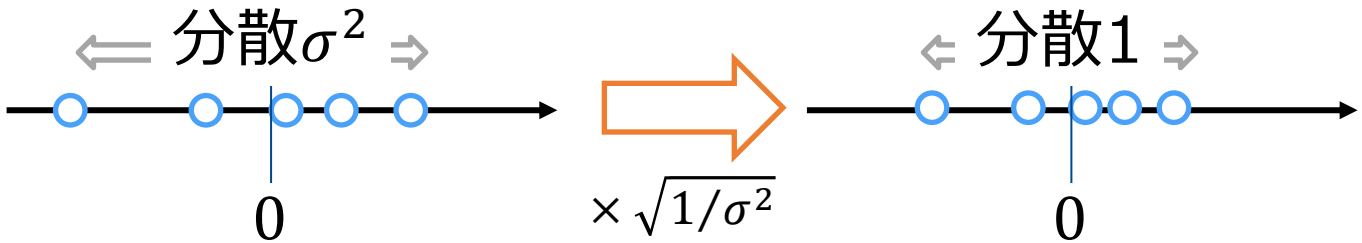


# 「平均ゼロ，分散を１」にするには？

- まず平均をゼロに
  - 全部の値から現在の平均値を引けば，平均はゼロになる




- 次に分散を１に！
  - 前に「値を $x_i$ から $\alpha x_i$ にすると，分散は $\alpha^2\sigma^2$ になる」と言いました
  - $\alpha^2\sigma^2 = 1$ にしたいということは，$x_i$ を $\alpha = \sqrt{1/\sigma^2}$ 倍すればOK!



「見せかけの数値」の距離を測るのは危険

実データ間の距離を測る際の留意点(2/3)

# ユークリッド距離が測れる＝差が定義できる

- アンケート結果やランキング(順位)を含むデータについて，ユークリッド距離を測るのは「本当は」間違い

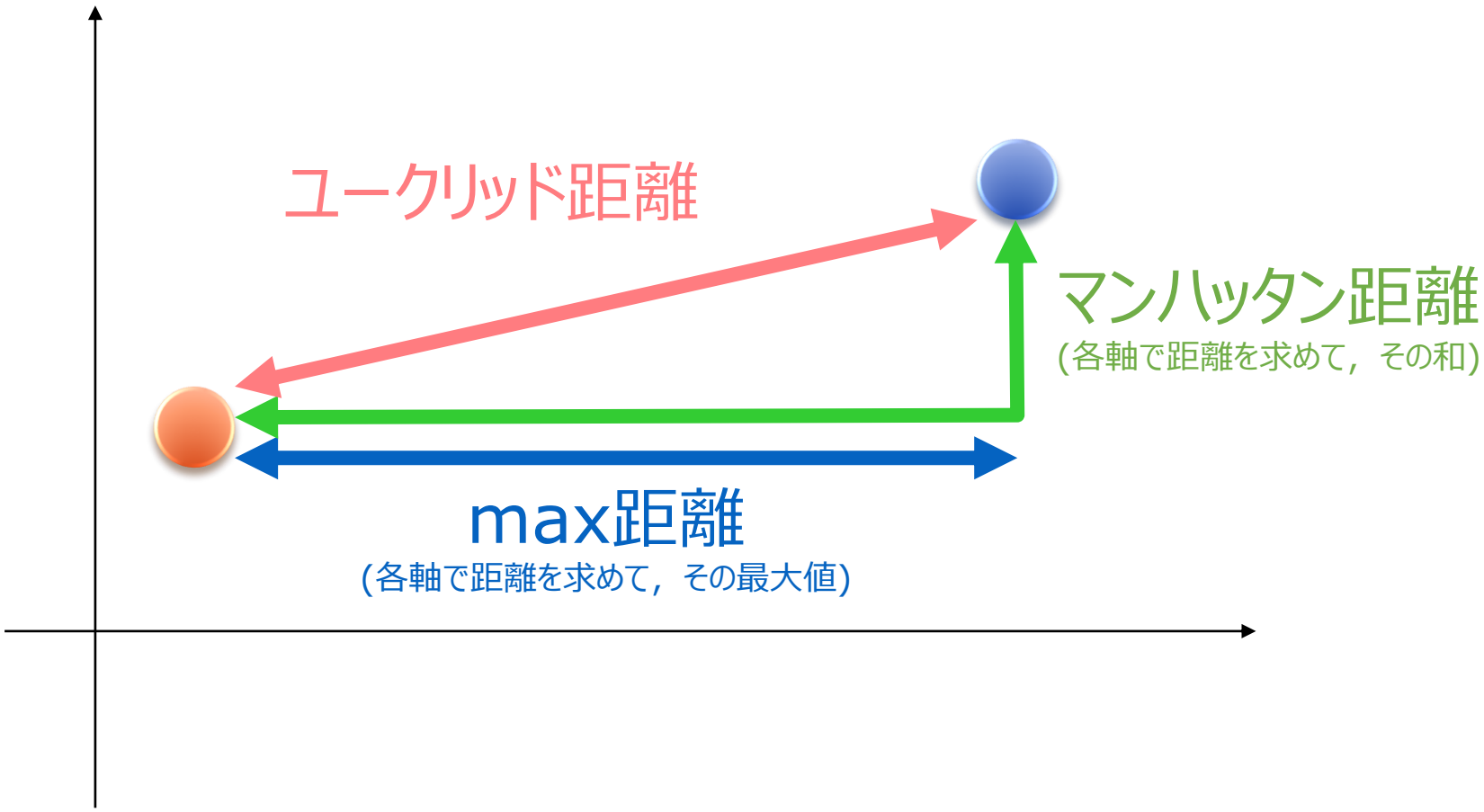
| | | 名称 | 可能な演算 | 主な代表値 | 主な事例 |
|---|---|---|---|---|---|
| 距離が測れる | 量的データ | 比率データ | ＋－×÷ | 各種平均 | 質量，長さ，年齢，時間，金額 |
| | | 間隔データ | ＋－ | 算術平均 | 温度（摂氏），知能指数 |
| 距離は測れない | 質的データ | 順位データ | ＞＝ | 中央値，最頻値 | 満足度，選好度，硬度 |
| | | カテゴリデータ | 度数カウント | 最頻値 | 電話番号，性別，血液型 |

※後述の離散距離なら，カテゴリデータの距離でも測れる

ユークリッド距離だけじゃない

## 実データ間の距離を測る際の留意点(3/3)

# ユークリッド距離だけじゃない：様々な距離



# マンハッタン距離？

- 斜めには行けない街での距離
  - 平安京距離
  - 平城京距離
  - 札幌距離

  と呼んでもよいはず

- 「市街地距離」と呼ばれることも

# max距離をいつ使う？

- 次の$d$次元データ間の距離を考えてみましょう

$$
d\text{個}\updownarrow \begin{pmatrix} 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} \qquad \begin{pmatrix} 2 \\ \vdots \\ 2 \\ 10 \\ 2 \\ \vdots \\ 2 \end{pmatrix}
$$

  - ユークリッド距離では，この差は小さい

- 「１要素でも大きく違ったら，それは結構違うのだ」としたい場合に
  - ただし1要素間でのみの評価になるので，全体的な差異は評価できない

# 式で書くと…. 実は統一的に書ける

$$L_p\text{距離} = \left( \sum_{i=1}^{d} \left| x_i - y_i \right|^p \right)^{1/p}$$

- マンハッタン距離 → $L_1$距離 (上の式において$p = 1$)
- ユークリッド距離 → $L_2$距離 (上の式において$p = 2$)
- max距離 → $L_\infty$距離 (上の式において$p = \infty$)

# 参考：$L_\infty$がなぜmax?

$$L_\infty = \left( \sum_{i=1}^{d} |x_i - y_i|^\infty \right)^{1/\infty}$$

$d$個の絶対値のうち，
一番大きいものが支配的

例($d = 3$)

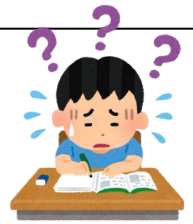
| | | 2乗 | 10乗 | 1000乗 |
|---|---|---|---|---|
| $\lvert x_1 - y_1 \rvert =$ | $10.1$ | $\rightarrow 102.01$ | $\rightarrow 1.1 \times 10^{10}$ | $\rightarrow 2.1 \times 10^{1004}$ |
| $\lvert x_2 - y_2 \rvert =$ | $10.2$ | $\rightarrow 104.03$ | $\rightarrow 1.2 \times 10^{10}$ | $\rightarrow 3.9 \times 10^{1008}$ |
| $\lvert x_3 - y_3 \rvert =$ | $10.3$ | $\rightarrow 106.09$ | $\rightarrow 1.3 \times 10^{10}$ | $\rightarrow 6.8 \times 10^{1012}$ |

あまり差はないが…

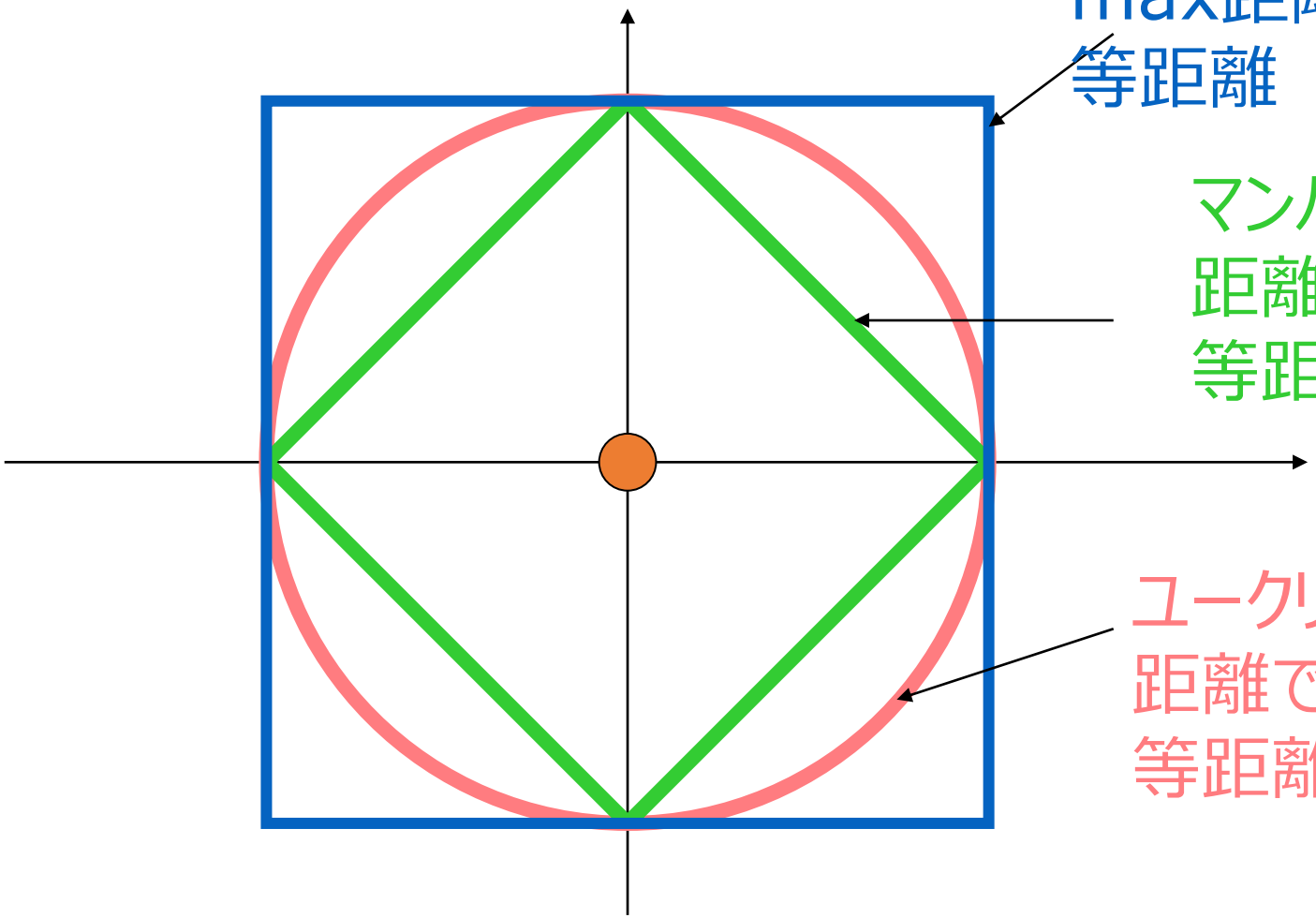
非常に大きな差となった！

最大以外のものは無視できる

よって $L_\infty = \left( \sum_{i=1}^{d} |x_i - y_i|^\infty \right)^{1/\infty} \sim \left( \max_{1 \le i \le d} |x_i - y_i|^\infty \right)^{1/\infty} = \max_{1 \le i \le d} |x_i - y_i|$



# 等距離面



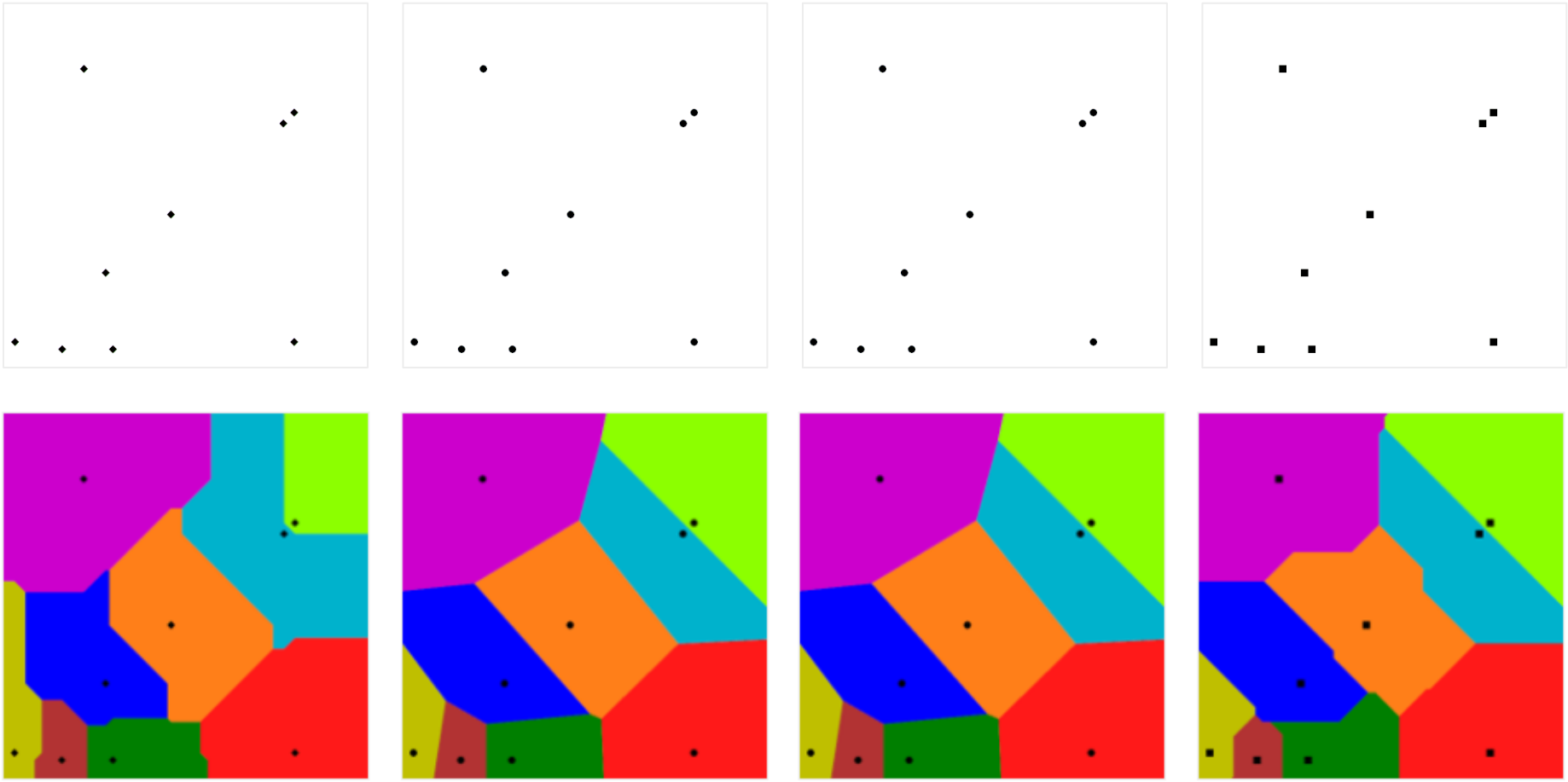
max距離で ● から
等距離

マンハッタン
距離で ● から
等距離

ユークリッド
距離で ● から
等距離

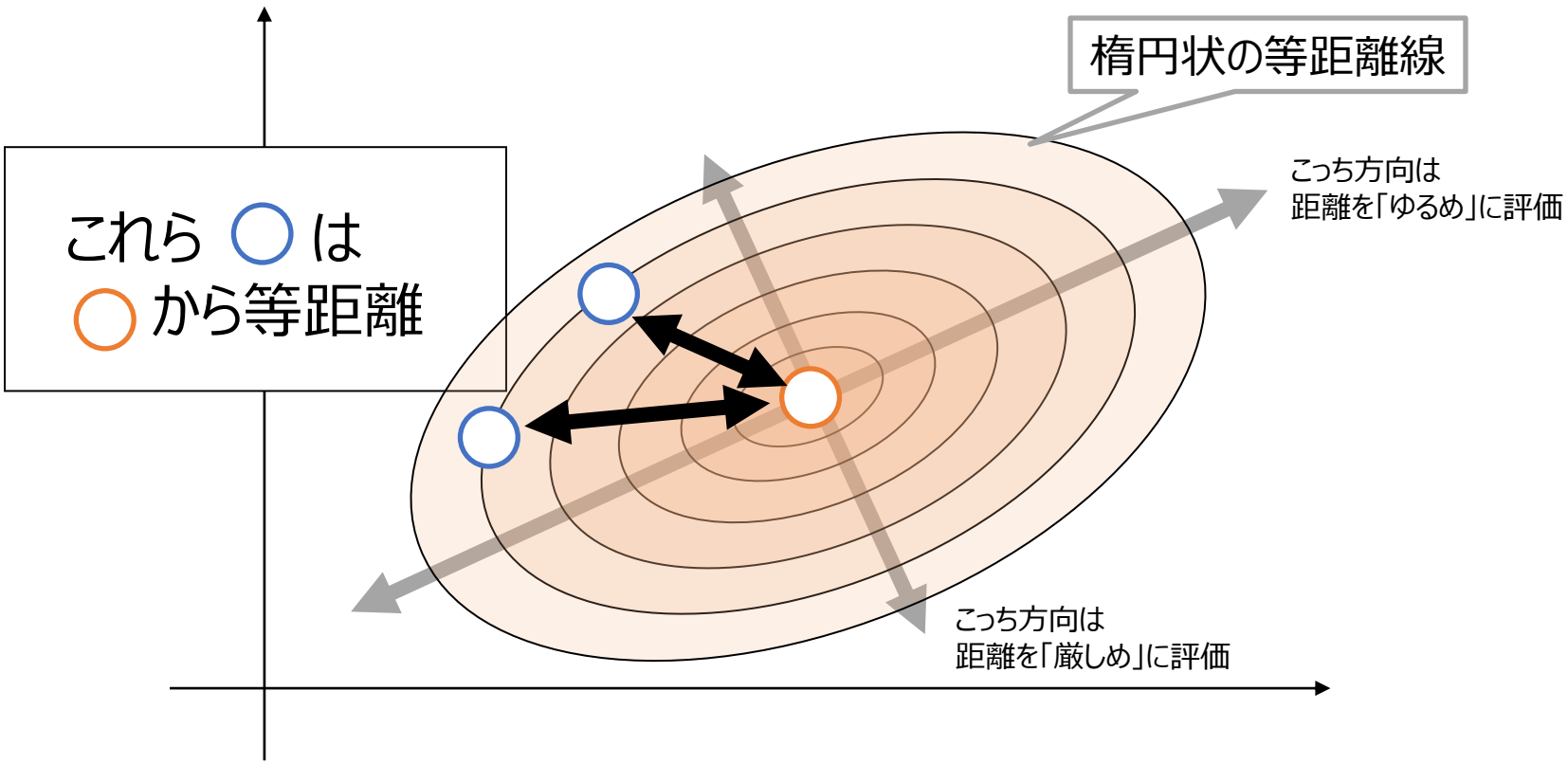
# ボロノイ図(縄張り図)を作ると…

$p = 1$ $p = 2$ $p = 3$ $p = \infty$



by [Jahobr](https://commons.wikimedia.org/wiki/File:Voronoi_growth_minkowski_p1_25.gif) https://commons.wikimedia.org/wiki/File:Voronoi_growth_minkowski_p1_25.gif

# マハラノビス距離 (Mahalanobis distance)




そのうち出てくる「正規分布」と密接に関係

# 離散距離

- 同じなら0，(少しでも)違うなら１


距離0


距離1

- どんなデータの距離でも測れます



2つのバスの番号(カテゴリデータ)

明らかに意味がない

ユークリッド距離$||2 - 5|| = 3$

離散距離＝1(違うから)

# ハミング距離 (Hamming distance)

- (長さの同じ)2系列間の距離
- 違う要素の数＝距離
- 例
  - 100101 ⇔ 110111 → 距離2
  - “Synchronize” ⇔ “Simchronise” → 距離3

# 編集距離 (edit distance)

- 2系列間の距離．「系列の長さが違っても大丈夫」がメリット
  - 置換，挿入，削除の最小回数
  - ハミング距離を一般化
  - Levenshtein距離とも

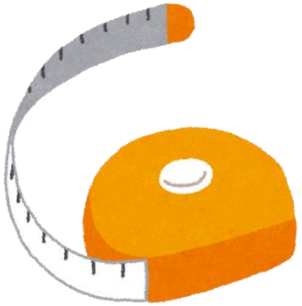
- 例
  - “This” ⇔ “These”
  - 置換1回(i⇔e)＋挿入1回(e)→ 距離 2
  - 削除1回(s)＋置換(i⇔e)1回＋挿入2回(se) → 距離 4
  - 削除2回(is)＋挿入3回(ese) → 距離 5
  - 削除4回(This)＋挿入5回(These) → 距離 9
  - ….

操作回数が最小で済む方法は自明ではない．何とかして見つける必要がある！
(操作回数の最小化問題)

こんな調子で距離には膨大な種類がある！

# そこで大事なことをもう一度！

- 距離は「データ解析の基本」である！
- 距離は1種類ではない！



- 距離が変われば，データ解析結果は「まるっきり」変わる
- 解析問題の性質に合致した「距離」を選ぶ必要がある
  - 逆に言うと，「自分で好きな距離を選べる」ともいえる

データ解析の基本中の基本が
自分のチョイスに任されているなんて…

数学は「だれが解いても同じ
答えが出る」と思っていたのに…

高校までの「ドリル」の世界の話．
数学の本質は「自由」！

# そう，この自由さがあるからこそ データ分析をしっかり学ぶ必要があるんです！

- 「これだけが唯一の方法」というものはない！
  - データ分析に「絶対的『真』」は存在しない

- だからこそ自分で考えないと！
  - 「このデータを，どういう手法で分析するのが妥当か？」

- だからこそ疑わないと！
  - 「この論文の分析結果は，どのような手法で得られたものか？」
  - 「また，なぜその手法をつかっているのか？ 」
  - 「都合の良い結果が得られるような，恣意的な手法はつかっていないか？」

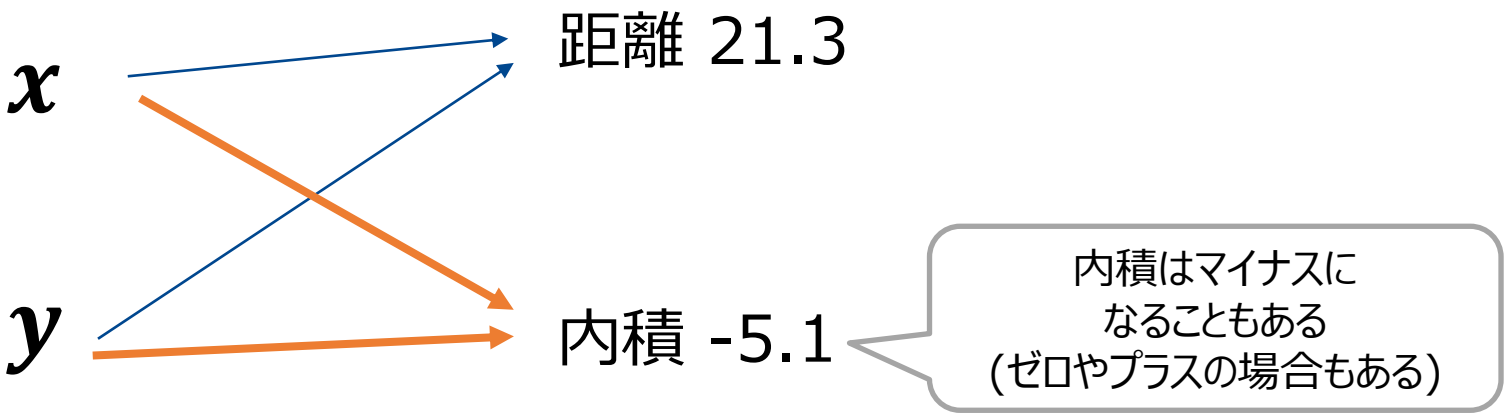
突然ですが…

# ベクトル間の内積

高校の時に習った内積．人生の役に立つことなんてないと思ってました？
データ解析やAIでは内積が超重要（皆さんの脳内でも常に内積計算が？）

# 内積？

- 距離と同様，2つのベクトルの関係を，1つの数字で表現
  - どっちが良いとか悪いとかいう話ではない．役割が違う




- なんでそんなもんが必要？
  - ベクトル(データ)間の類似度に使える
    - 類似度は「似てる具合」．距離は「似てない具合」．
  - そのうち非常に重要になってくる
    - 主成分分析とかディープラーニング(深層学習)とかはバリバリ

# 内積の計算法

内積の書き方4種(どれも同じ)

$$\boldsymbol{x}\cdot\boldsymbol{y}$$
$$(\boldsymbol{x},\boldsymbol{y})$$
$$\langle\boldsymbol{x},\boldsymbol{y}\rangle$$
$$\boldsymbol{x}^T\boldsymbol{y}$$

- 習うより慣れよう．こんな感じ．

$$\boldsymbol{x}=\begin{pmatrix}3\\5\end{pmatrix},\boldsymbol{y}=\begin{pmatrix}6\\1\end{pmatrix}\text{の内積}\rightarrow \boldsymbol{x}\cdot\boldsymbol{y}=3\times6+5\times1=23$$

- 要するに，「要素どうしの積をとって，全部足す」(小学生でも計算できる)
  - その原理で，何次元ベクトルでも計算可能

$$\begin{pmatrix}3\\5\end{pmatrix}\text{と}\begin{pmatrix}6\\1\end{pmatrix}\text{の内積}\Rightarrow\left.\begin{array}{l}\begin{pmatrix}3\\5\end{pmatrix}\times\begin{pmatrix}6\\1\end{pmatrix}\begin{array}{l}=18\\=5\end{array}\end{array}\right\}18+5=23$$

$$\begin{pmatrix}3\\5\\2\end{pmatrix}\text{と}\begin{pmatrix}6\\1\\2\end{pmatrix}\text{の内積}\Rightarrow\left.\begin{array}{l}\begin{pmatrix}3\\5\\2\end{pmatrix}\times\begin{pmatrix}6\\1\\2\end{pmatrix}\begin{array}{l}=18\\=5\\=4\end{array}\end{array}\right\}18+5+4=27$$

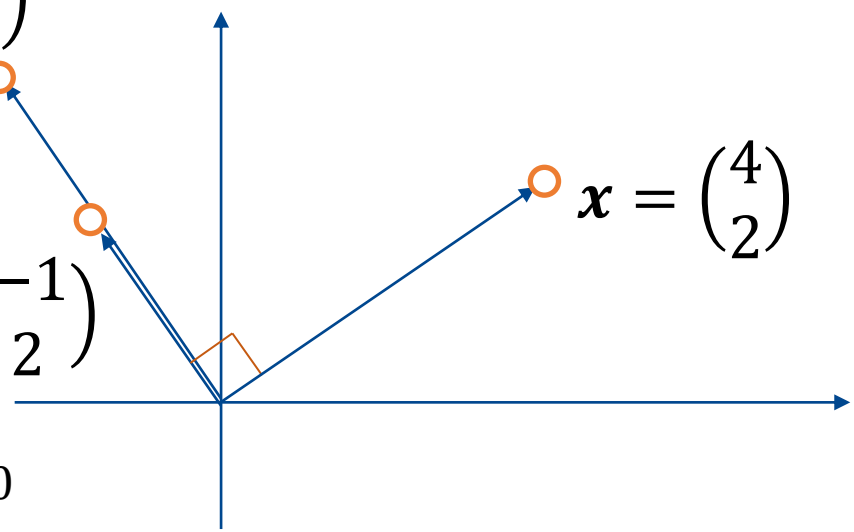
※この調子で，4次元でも，100万次元でも可能

# 内積の性質：<br>ベクトルのなす角が90度のときに，内積はゼロ！

- なんと美しい性質！

$$\boldsymbol{y} = \begin{pmatrix} -2 \\ 4 \end{pmatrix}$$

$$\boldsymbol{x} \cdot \boldsymbol{y} = 4 \times (-2) + 2 \times 4 = 0$$



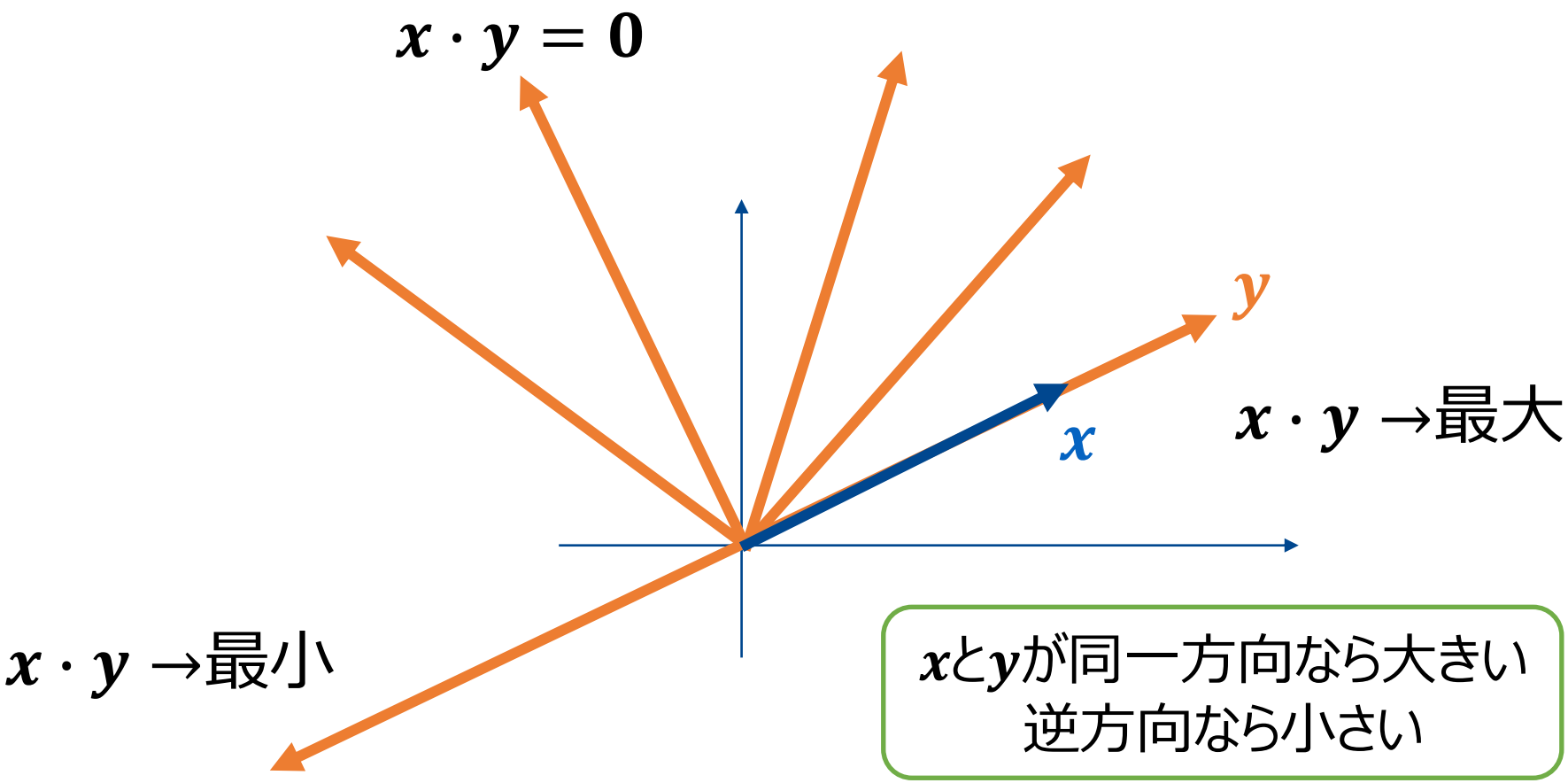

$$\boldsymbol{y} = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

$$\boldsymbol{x} \cdot \boldsymbol{y} = 4 \times (-1) + 2 \times 2 = 0$$

2次元以外でも
成り立つ…

# 内積の性質：ベクトルのなす角 $\theta$ が変わると内積も変わる

- 一定の大きさのベクトル$\boldsymbol{y}$を回転させながら$\boldsymbol{x}$と内積を取ると…

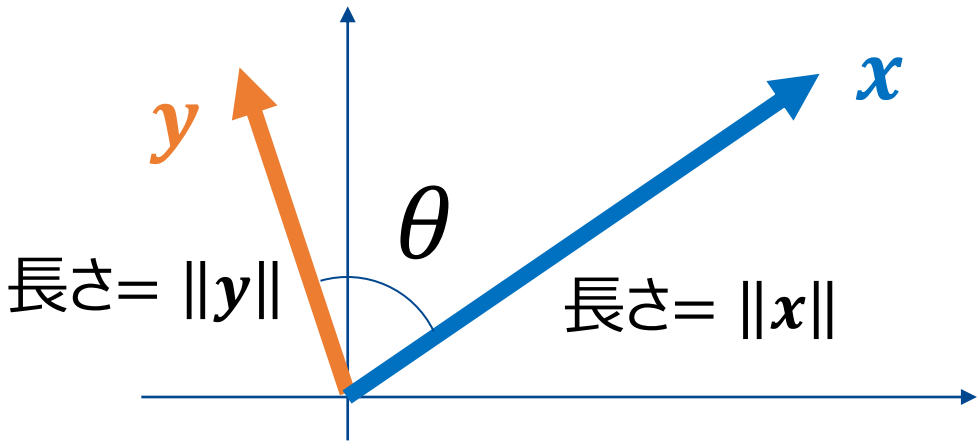



$\boldsymbol{x}$と$\boldsymbol{y}$が同一方向なら大きい
逆方向なら小さい

# 内積の性質：だったら，ベクトルのなす角$\theta$を使って書けるんじゃ？

- $\cos\theta$ を使って書けます

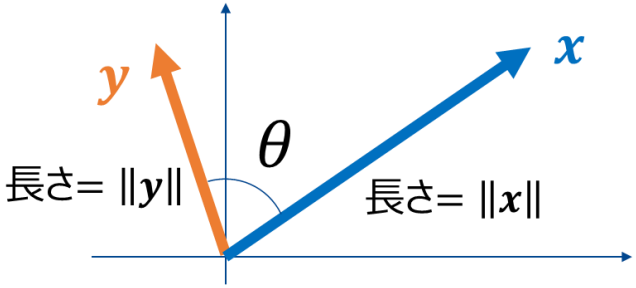
$$\boldsymbol{x}\cdot\boldsymbol{y} = \|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos\theta$$




- なんか不思議ですよね
  - 要素を掛け算して足したもの ＝ ベクトルの長さとその間の角度で計算されたもの

# 内積の性質：まとめると

$$\boldsymbol{x}\cdot\boldsymbol{y} = \|\boldsymbol{x}\|\|\boldsymbol{y}\| \cos\theta$$

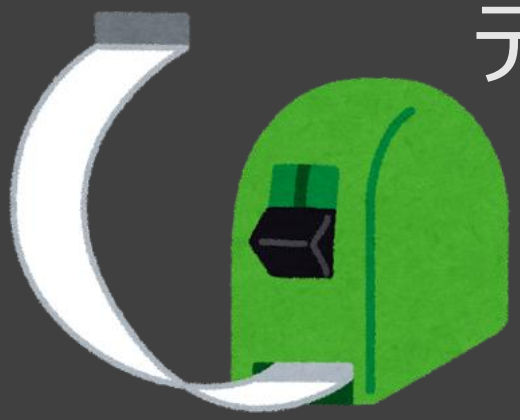


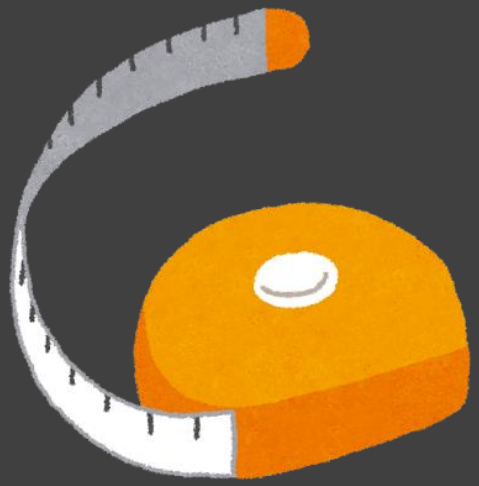

- ベクトルが長いほど(＝$\|\boldsymbol{x}\|$と$\|\boldsymbol{y}\|$が大きいほど)内積値は大きくなりやすい

- $-1 \leq \cos\theta \leq 1$ だから，内積値は…
  - $\theta = 0^\circ$の時に最大となって　$\|\boldsymbol{x}\|\|\boldsymbol{y}\|$ ，
  - $\theta = 180^\circ$の時に最小になって $-\|\boldsymbol{x}\|\|\boldsymbol{y}\|$

- そして $\cos\theta = 0$の時，すなわち$\theta = 90^\circ$ のとき，内積値はゼロ，

# データ間の類似度

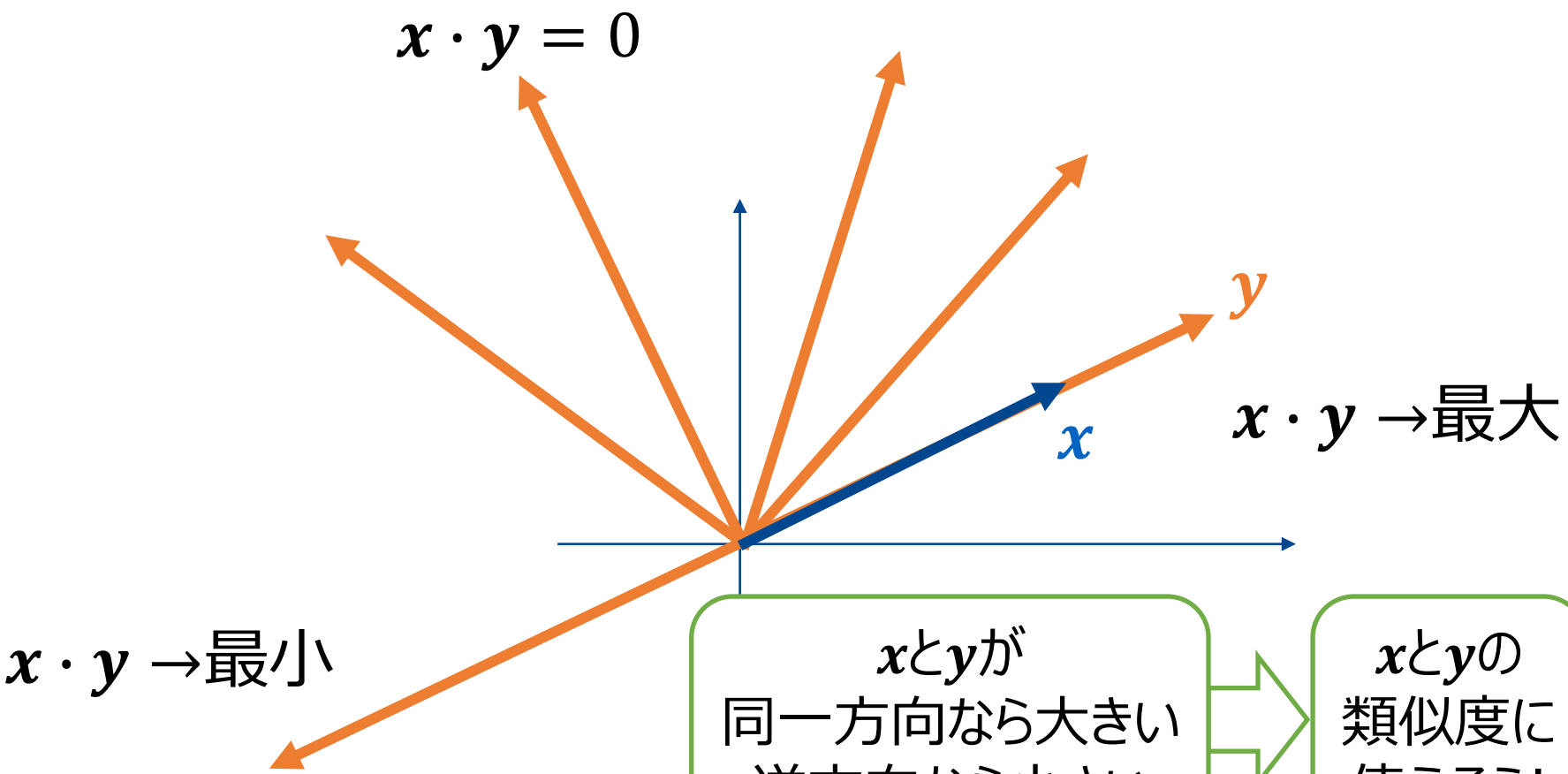




距離と逆で，似てると大きくなる！
内積をつかって計算できる！

# というわけで：
# ベクトルのなす角 $\theta$ が変わると内積も変わる

- 一定の大きさのベクトル$\boldsymbol{y}$を回転させながら$\boldsymbol{x}$と内積を取ると…



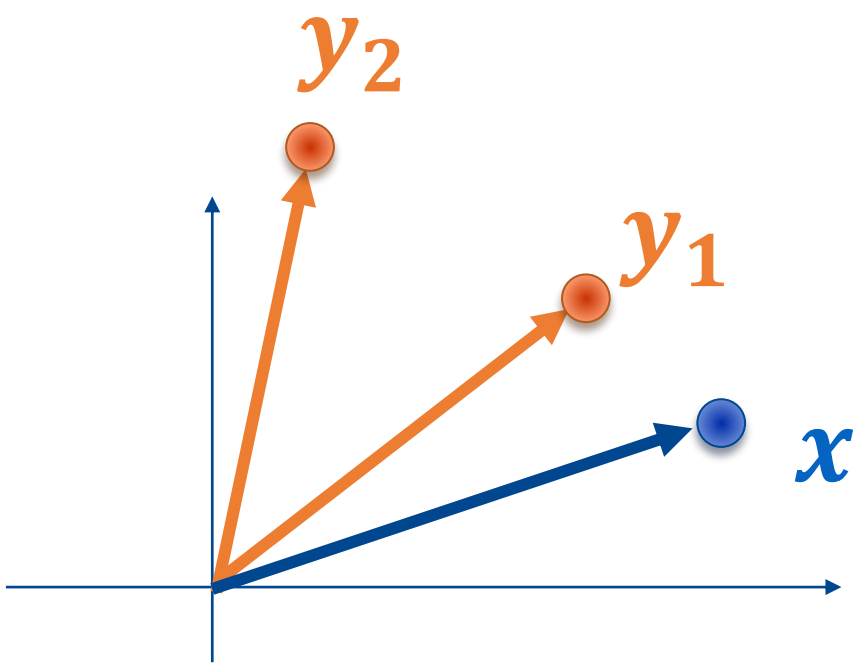

$\boldsymbol{x}$と$\boldsymbol{y}$が
同一方向なら大きい
逆方向なら小さい

⇒

$\boldsymbol{x}$と$\boldsymbol{y}$の
類似度に
使えそう!

# 内積は類似度として使える？（1/2）

- 使えるかも！
- 次の図なら，$\boldsymbol{x} \cdot \boldsymbol{y}_1 > \boldsymbol{x} \cdot \boldsymbol{y}_2$　（$\boldsymbol{x}$にとって，$\boldsymbol{y}_1$のほうが似ている(近い)）

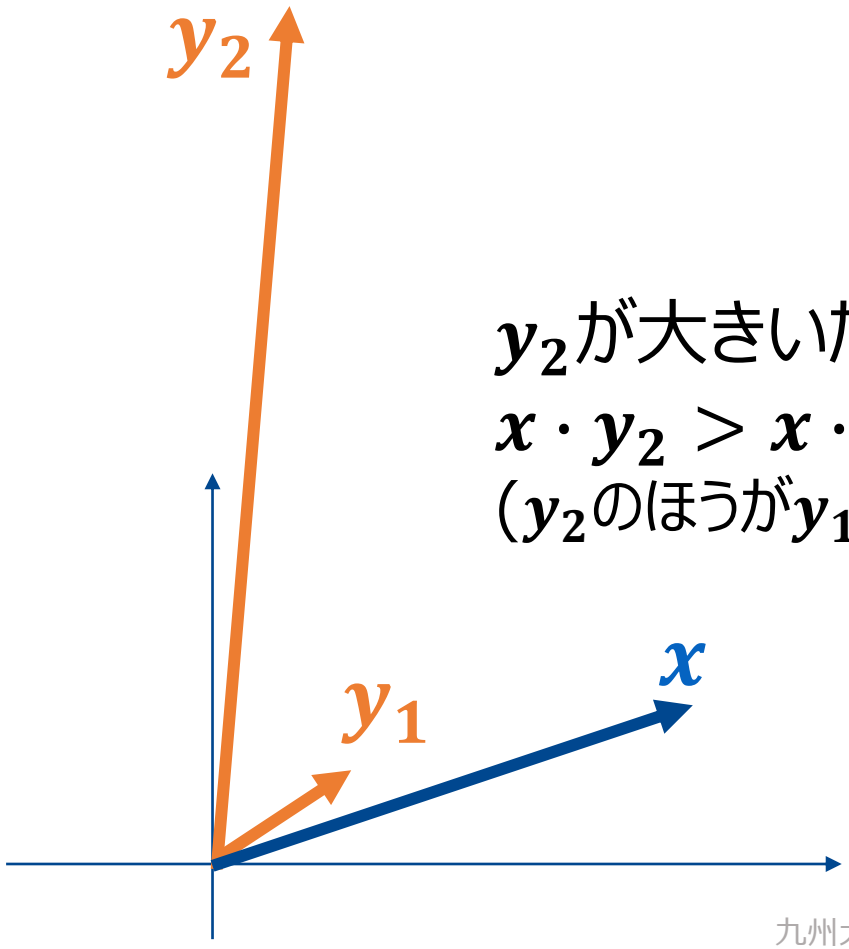



※実は機械学習やAIでは内積を類似度として使います

# 内積は類似度として使える？(2/2)

- でも内積値はベクトルの大きさに依存するので…

$$\boldsymbol{x}\cdot\boldsymbol{y} = \|\boldsymbol{x}\|\|\boldsymbol{y}\|\cos\theta$$




$\boldsymbol{y}_2$が大きいために
$\boldsymbol{x}\cdot\boldsymbol{y}_2 > \boldsymbol{x}\cdot\boldsymbol{y}_1$ となってしまう
($\boldsymbol{y}_2$のほうが$\boldsymbol{y}_1$よりも$\boldsymbol{x}$に類似!?)

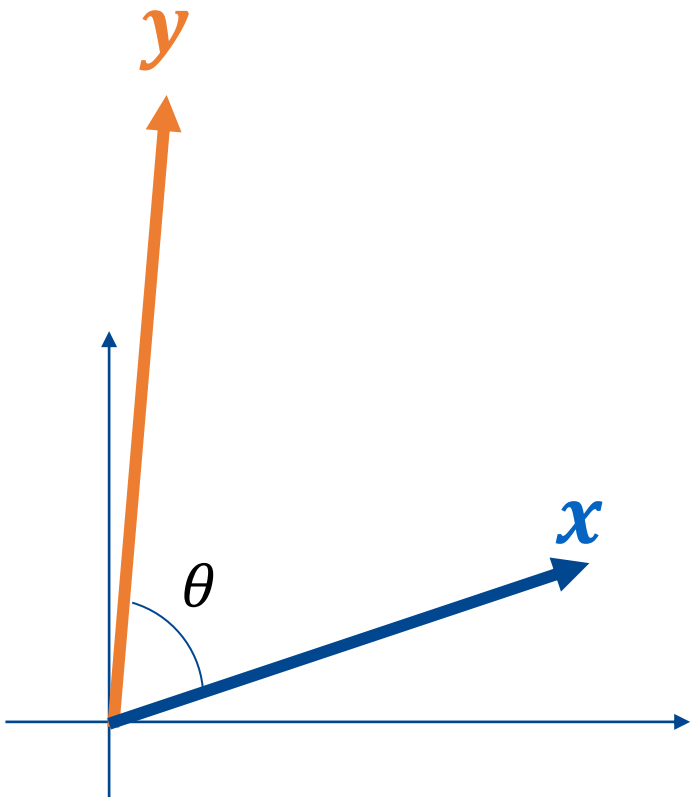
# 最も代表的な類似度：正規化相関

- ベクトルの大きさの影響が問題なら，
- 大きさに影響されないcos θ を使えばいいのでは？

$$\boldsymbol{x} \cdot \boldsymbol{y} = \|\boldsymbol{x}\| \|\boldsymbol{y}\| \cos\theta$$

$$\cos\theta = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}$$

結構簡単に求まる…



# 【付録１】
# 距離とは何か？

数学の本質はその自由さにあり！

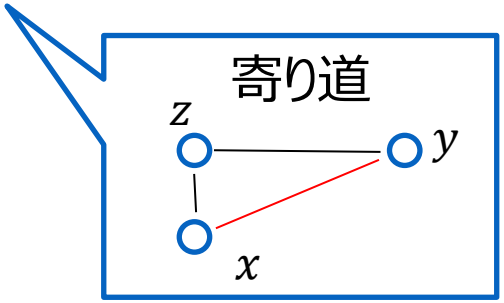
答えが唯一の「ドリル」が数学ではない.

きちんと定義さえ守れば何でもあり！それが数学の本当の姿

# そもそも距離ってなんだ？(1/2)

- 数学的には，次の３条件を満たす$d(x,y)$を$x,y$の「距離」と呼ぶ
  - 非退化性(同じものだけ距離がゼロ)： $x = y \Leftrightarrow d(x,y) = 0$
  - 対称性(「$x$から$y$へ」と「$y$から$x$へ」の距離は同じ): $d(x,y) = d(y,x)$
  - 三角不等式(寄り道したら遠くなる)： $d(x,z) + d(z,y) \geqq d(x,y)$

↑「距離の公理」と呼ばれる
(公理＝決めごと)




- 条件を満たすなら，何でも「距離」
  - 山本君が「山本距離」を勝手に作ってもOK
  - ルールさえ満たせば，何作ってもOK!

数学の本質は
その自由さにある
The essence of mathematics
is its freedom.

G. Cantor (1845-1918)

# そもそも距離ってなんだ？(2/2)

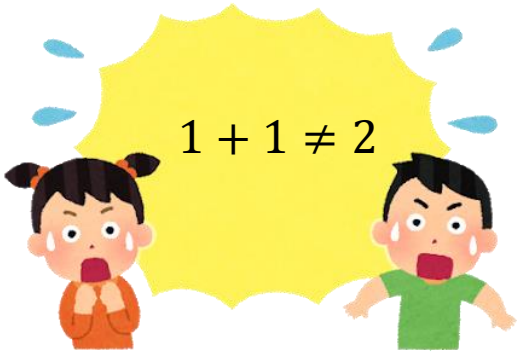
- 実用上は上記条件を満たさない$d(x, y)$を使うことがある
  - 正確には「擬距離」(pseudo-distance)と呼ばれる
  - 対称性を満たさない場合が多い
  - Ex. 2地点間の距離を，「所要時間」で測ると…

登りと降りで
距離が違う!?

# 参考：数学が自由さを見せる例
# ブール代数(1+1=1の世界)

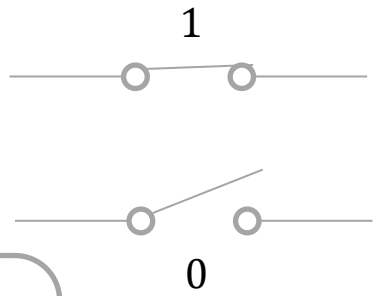
- 0と1しかない世界で定義された数学
- 和と積しかない
- ルール
  - $0 + 0 = 0$
  - $0 \cdot 0 = 0$
  - $1 + 0 = 1$
  - $1 \cdot 0 = 0$
  - $1 + 1 = 1$
  - $1 \cdot 1 = 1$

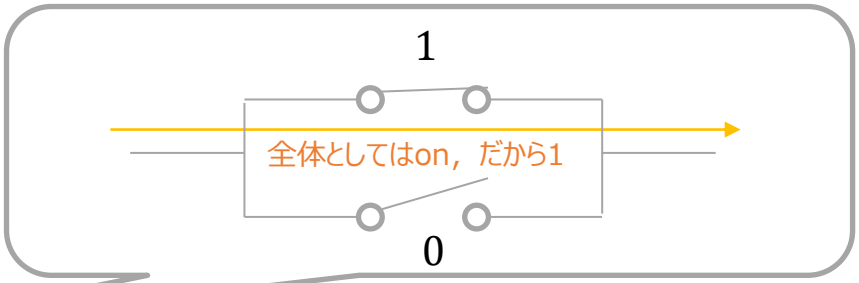
小学校1年で習った$1 + 1 = 2$という常識が
通用しない世界！

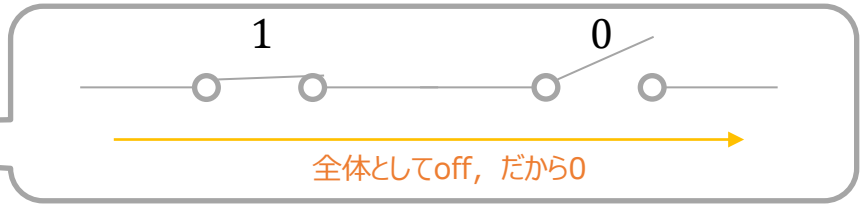


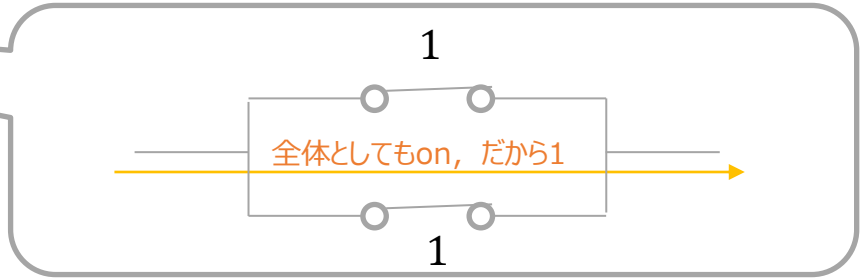
# 参考：数学が自由さを見せる例 ブール代数(1+1=1の世界)

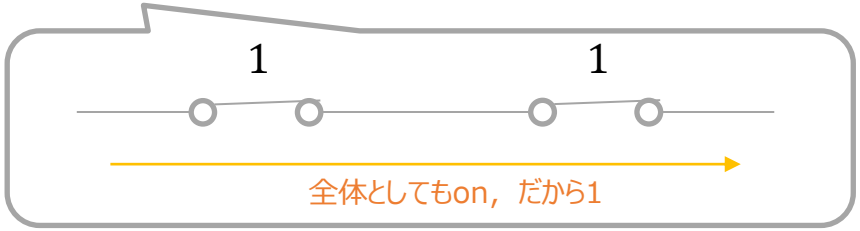
- 1=1個ではなく「1=on，0=off」と考える
- 和＝並列つなぎ，積＝直列つなぎ
- ルール
  - $0+0=0$
  - $0\cdot 0=0$
  - $1+0=1$
  - $1\cdot 0=0$
  - $1+1=1$
  - $1\cdot 1=1$



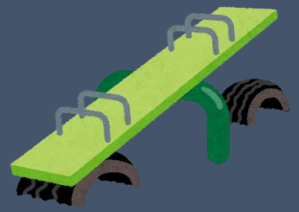




要するに{0,1}や{+,・}の意味が，小学校の時と違う！

# 【付録２】
# ユークリッド距離と内積の関係

実は関係してます



# ユークリッド距離と内積の関係

- ユークリッド距離の計算を展開すると内積が出てくる

$$\underbrace{\|\boldsymbol{x}-\boldsymbol{y}\|^2}_{\text{ユークリッド距離(の二乗)}} = (\boldsymbol{x}-\boldsymbol{y})^2 = \|\boldsymbol{x}\|^2 - 2\underbrace{\boldsymbol{x}\cdot\boldsymbol{y}}_{\text{内積}} + \|\boldsymbol{y}\|^2$$

- なので，両者は無関係ではない！
- 特に$\|\boldsymbol{x}\|$や$\|\boldsymbol{y}\|$が一定なら，
  ユークリッド距離大→内積小