

データサイエンス概論I & II データサイエンス総論I & II

データのベクトル表現と集合

九州大学 数理・データサイエンス教育研究センター

データのベクトル表現

そんな高尚な話ではありません.
皆さんも無意識にベクトルを考えている!?

「ベクトル = 数字の組」

- A君： 体重62kg, 身長173cm
 - A君の体格データは, 2つの数字の組で表される

(62, 173)



- これがわかれば, もう大丈夫 !
 - 第一段階突破

「ベクトル = 数字の組」

- B君: 体重57kg, 身長164cm
 - B君の体格データも, 2つの数字の組で表される

(57, 164)

「ベクトル = 数字の組」

- A君: 体重62kg, 身長173cm, 腹囲78cm
 - A君の体格データは, 3つの数字の組でも表される

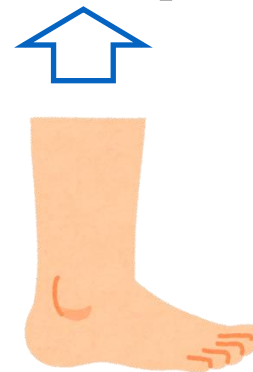
(62, 173, 78)



「ベクトル = 数字の組」

- A君: 体重62kg, 身長173cm, 腹囲78cm, 靴のサイズ26cm
 - A君の体格データは, 4つの数字の組でも表される

(62, 173, 78, 26)

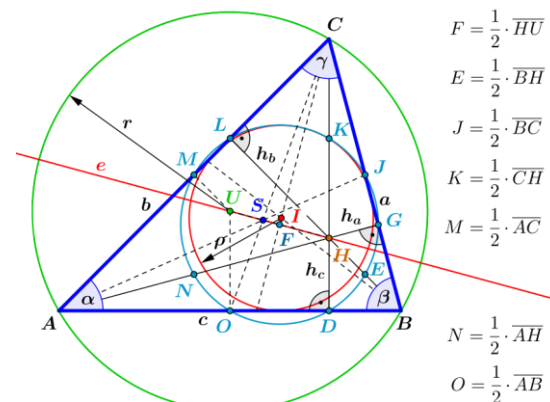


なぜいきなり「ベクトル」とやらを学ぶ？ データサイエンスとどんな関係が？







- 実データを扱う研究では、実は頻繁に「ベクトル」が出てきます
 - 例えばデータが「血糖値と体重」のような数字の組なら、もうそれは「ベクトル」
 - 画像や時系列データも、実は「ベクトル」とみなせます
 - アンケートデータでも、いろいろな形で「ベクトル化」できます
- 実際、主成分分析や因子分析など多くの有名な解析手法のターゲットは、「ベクトル」なんです
- というわけで、高校・大学学部時代に
苦い思いをした皆さんも、データ解析の
視点から、ベクトルについて再度考えて
みましょう


↓ こういうのはとりあえず出てこない



Petrus3743@Wikipedia Commons

(あとでも言いますが)
表にもベクトルが潜んでいる…

	 1	 2	 3	...	 N
体重	62	57	65	...	75
身長	173	164	171	...	164



各人の体格を1つのベクトルで表している

「次元」とは？

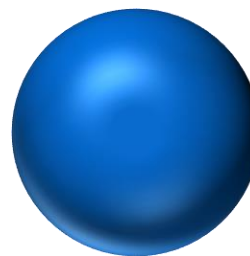
- 組み合わせた数字の数，です
 - $(62, 173) \rightarrow 2\text{次元ベクトル}$
 - $(62, 173, 78) \rightarrow 3\text{次元ベクトル}$
 - $(62, 173, 78, 26) \rightarrow 4\text{次元ベクトル}$
 - $(62, 173, 78, 26, \dots, 9) \rightarrow 200\text{億次元ベクトル}$
- 200億個の数字
- 特にそれ以上の意味はない
 - 以下のような物理的な意味づけは，**とりあえず忘れましょう**



1次元



2次元



3次元

?

4次元

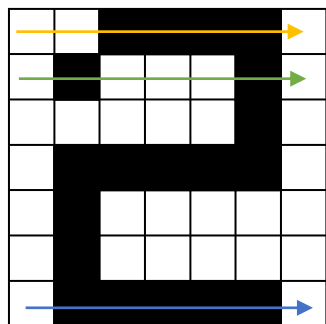
おまけ：ちょっと考えてみよう

- あなた自身を表すとすれば、何次元ベクトルになるでしょう？
- 例えば30問の性格診断に答えたら、30次元ベクトルが得られる
 - これも「あなた」を表現するベクトルですね
 - でもそれで十分なのかなあ…
- 人間の細胞の数は37兆個らしいです。もし、各細胞の性質(場所とか大きさとか？)が100個の数字で表されたとすれば、3700兆？
 - そんなに必要かなあ…
- 答えはないです
 - でも色々考えてみると面白い
 - 毎日ちょっとずつ違うベクトルかもしれませんね

そんな高い次元なんて必要なの？

→ **画像**をベクトルで表現してみる (1/2)

● 2値画像の例

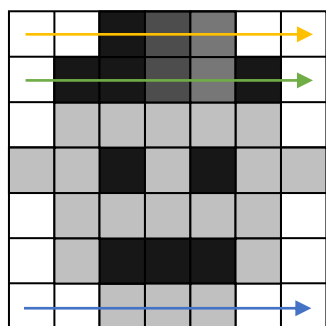


7×7画素

$$\Rightarrow (1, 1, 0, 0, 0, 0, 1, 1, 0, \dots, 0, 1)$$

49次元ベクトル

● グレースケール画像の例



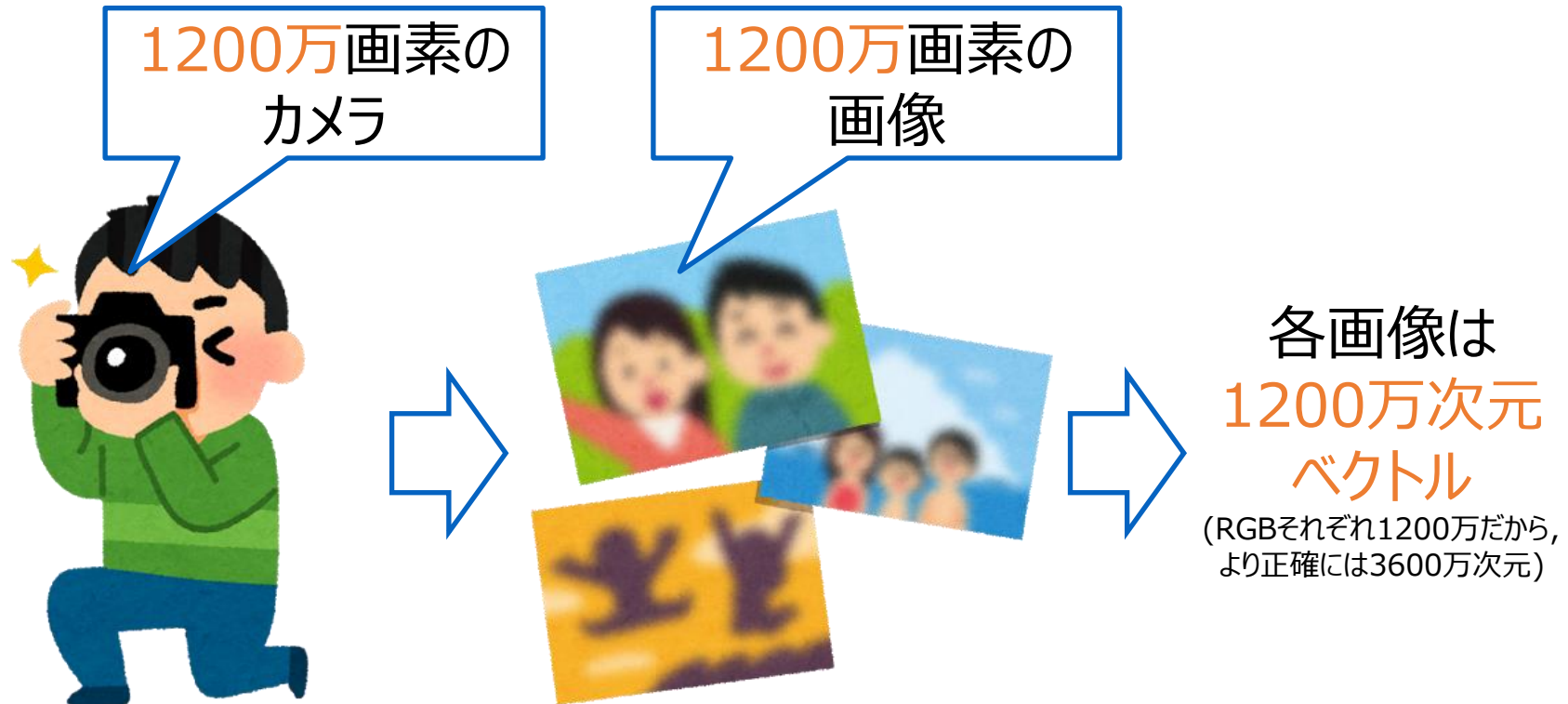
7×7画素

$$\Rightarrow (255, 245, 10, 35, 92, 231, 254, \dots, 249)$$

49次元ベクトル

そんな高い次元なんて必要なの？

→ 画像をベクトルで表現してみる (2/2)

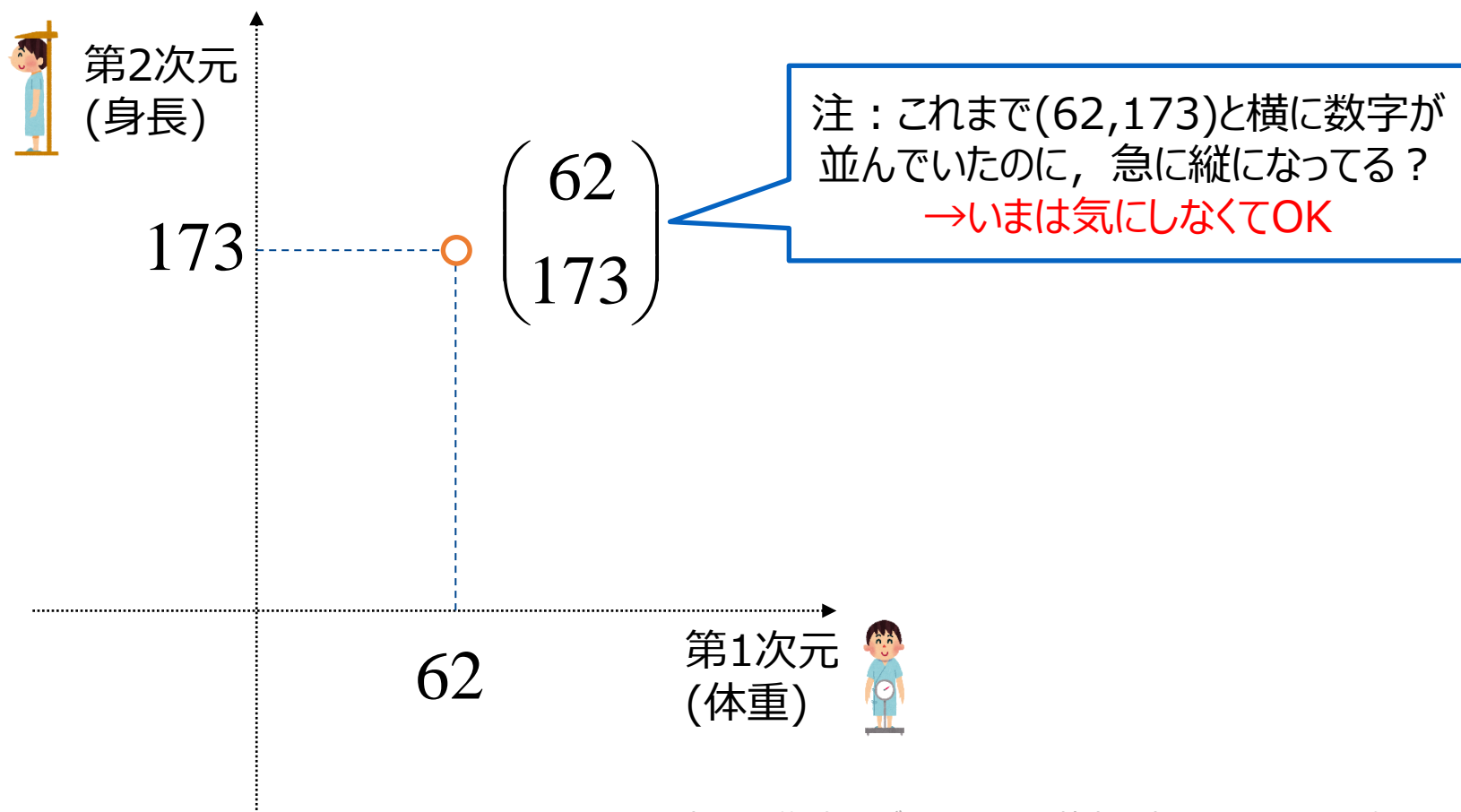


- 皆さんのスマホ・デジカメ・コンピュータは、いつも超高次元ベクトルを扱っている
 - シャッター押した瞬間に1200万次元ベクトルが一つ生まれている

ベクトルと座標系：

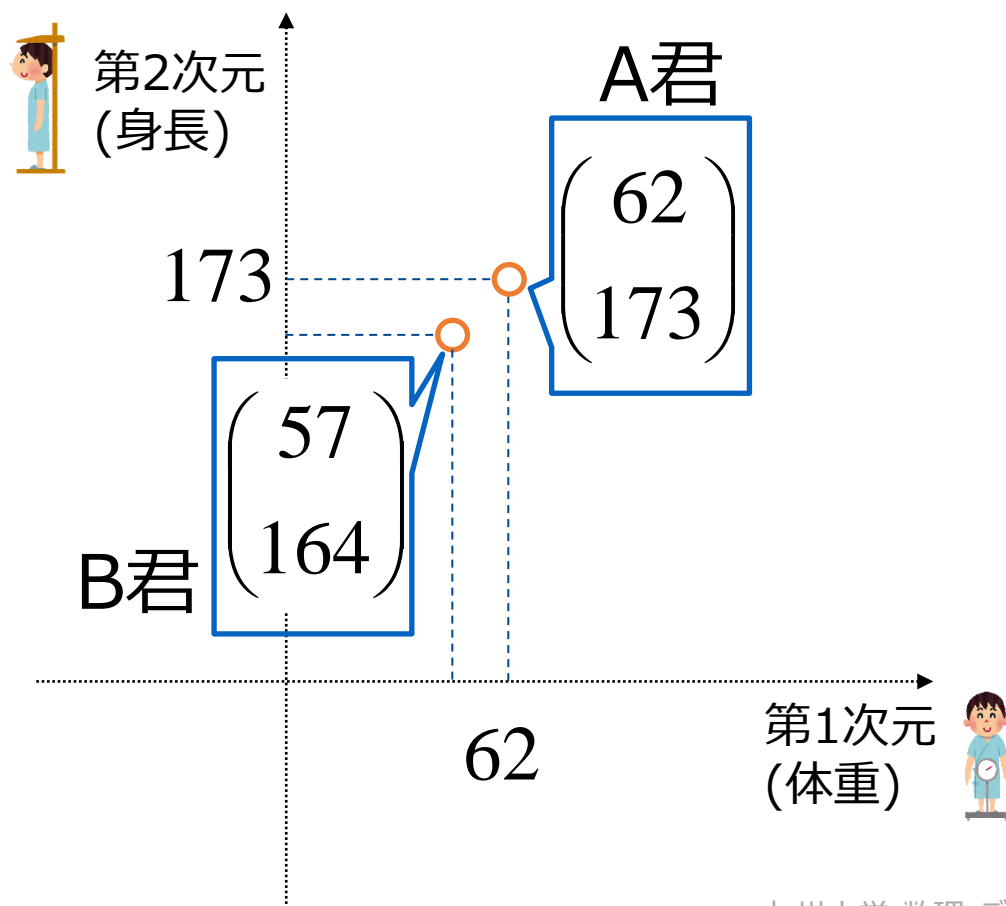
1つの2次元ベクトルは，2本の座標軸を使った平面上で表現できる

- 1 データ(1ベクトル)が1点で表される

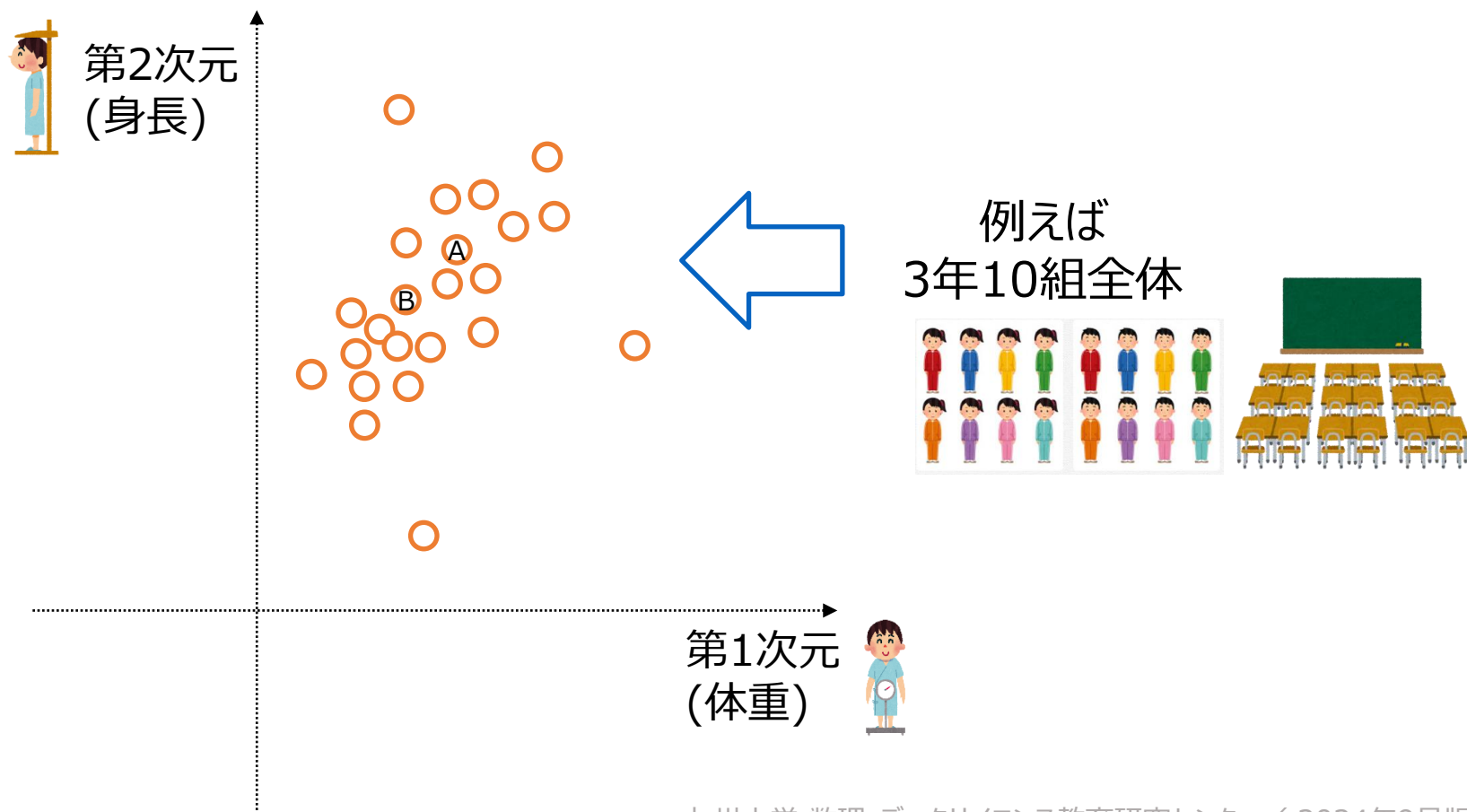


ベクトルと座標系：

「複数の点で複数のデータを同時に図示」もできます

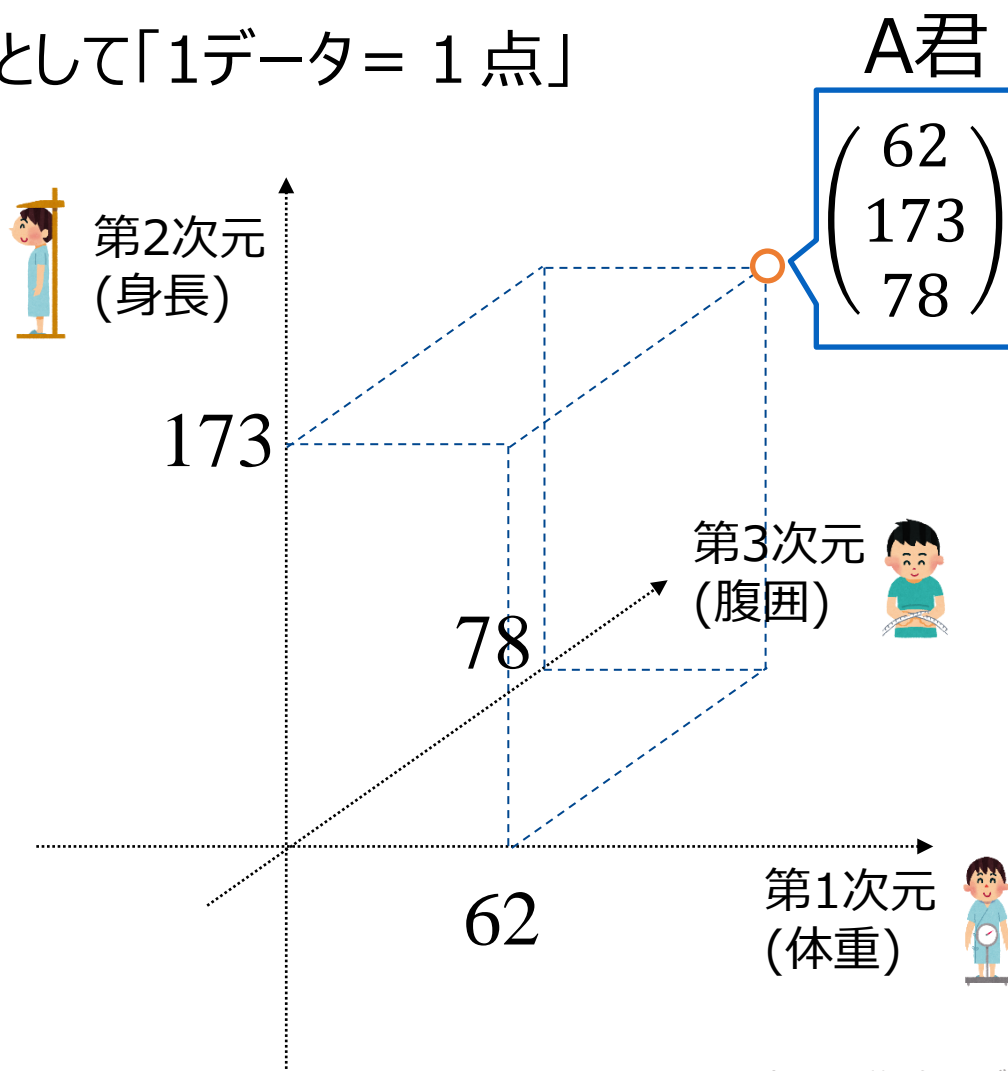


ベクトルと座標系： たくさんデータがあると「点群」になる

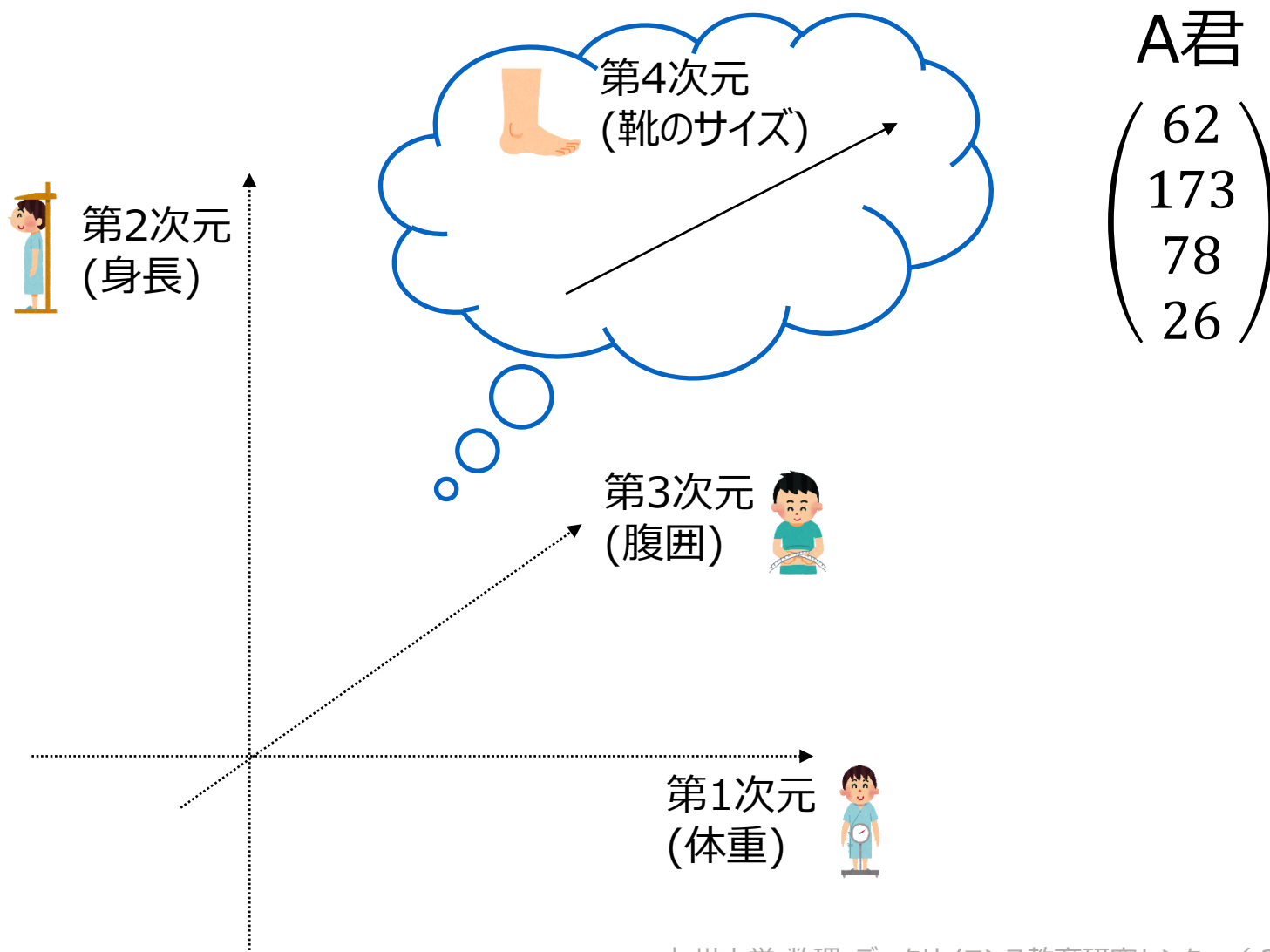


ベクトルと座標系： 3次元でもOK

- 依然として「1データ = 1点」



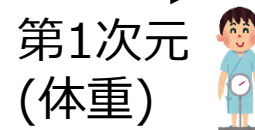
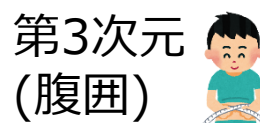
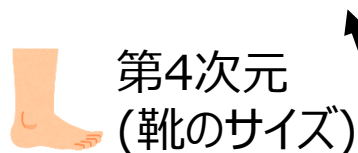
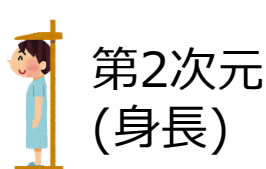
ベクトルと座標系： さて4次元. どこに座標軸を書けば...？



ベクトルと座標系： そんなときは、えいやっと「イイカゲン表現」

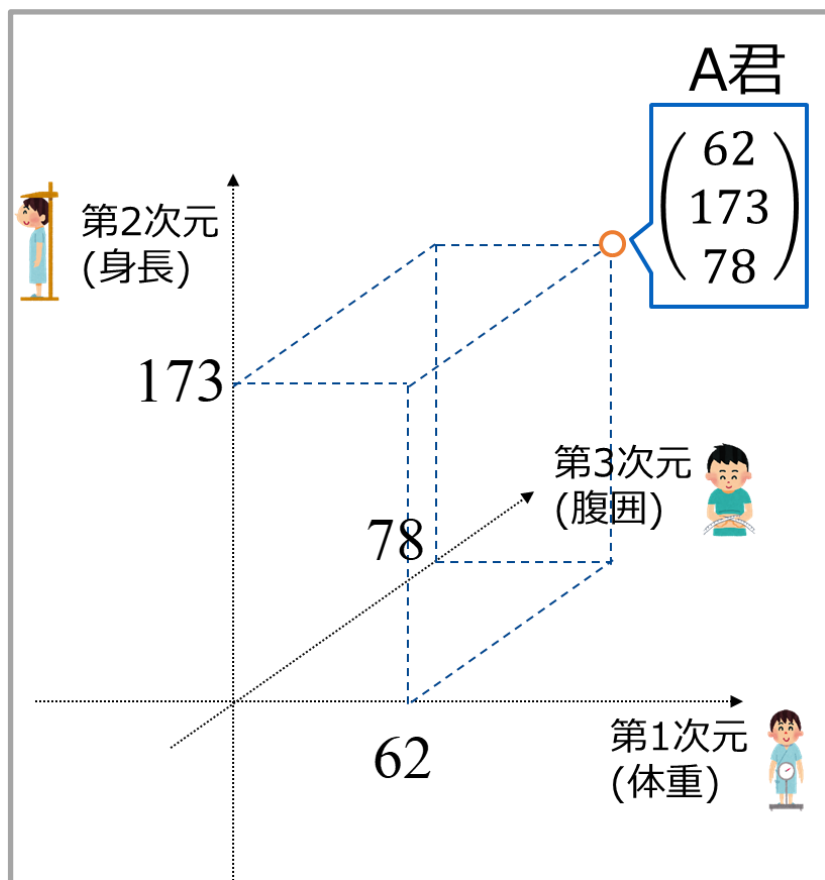
- この寛容さ(=抽象的思考)がデータサイエンスではすごく大事

要は「3次元も4次元も、
次元が1つ増えただけで、
大して変わらない」という
気持ちで行こう！



どこ向いているのかよく
わからないが、とにかく
第4の座標軸がある

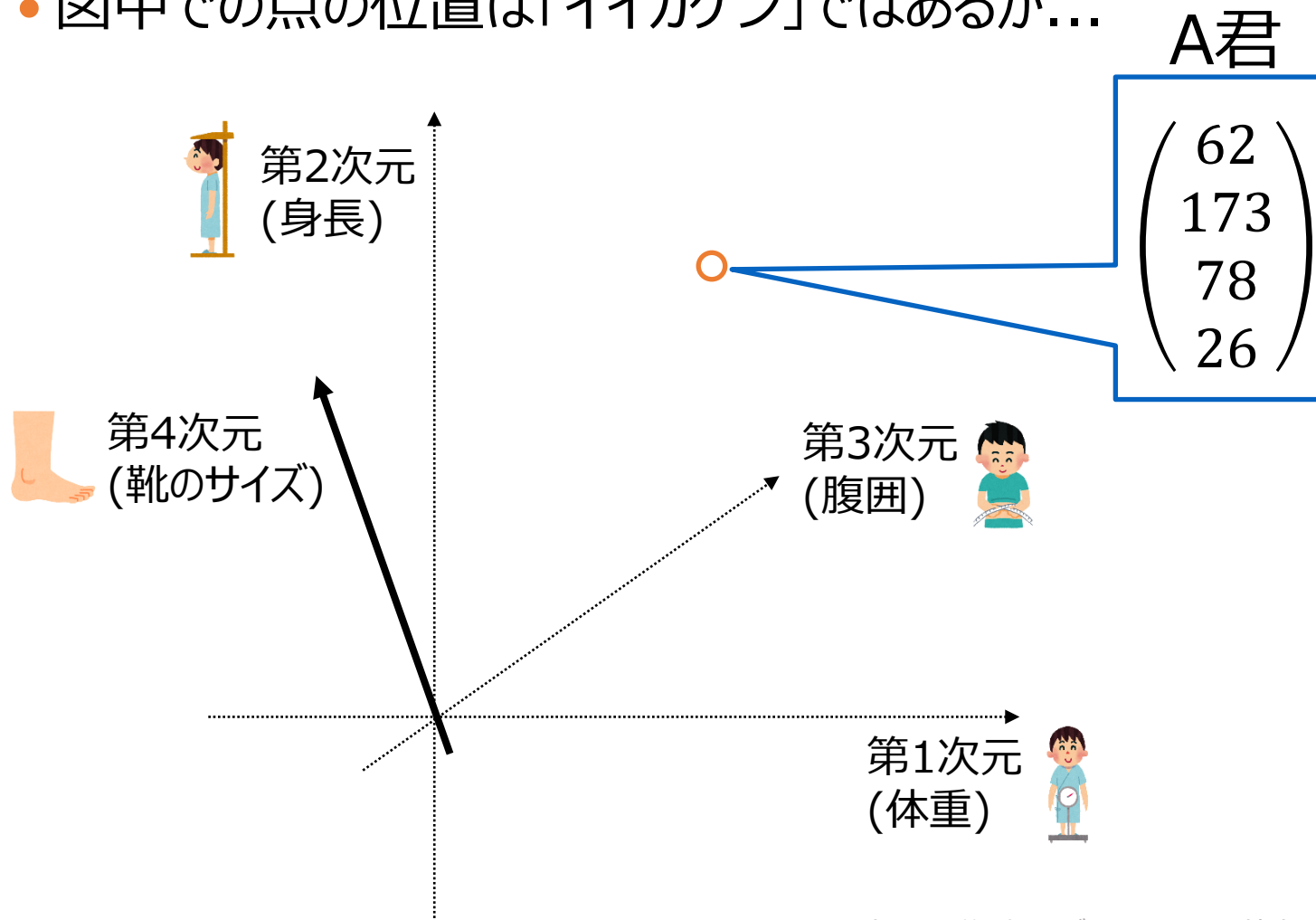
余談: そもそも3次元だって怪しいわけで...
だから4次元も許してやってください



3次元を無理に2次元表示している
わけですから、点の位置は怪しい。
(そもそも座標軸が90度で交わってないし)

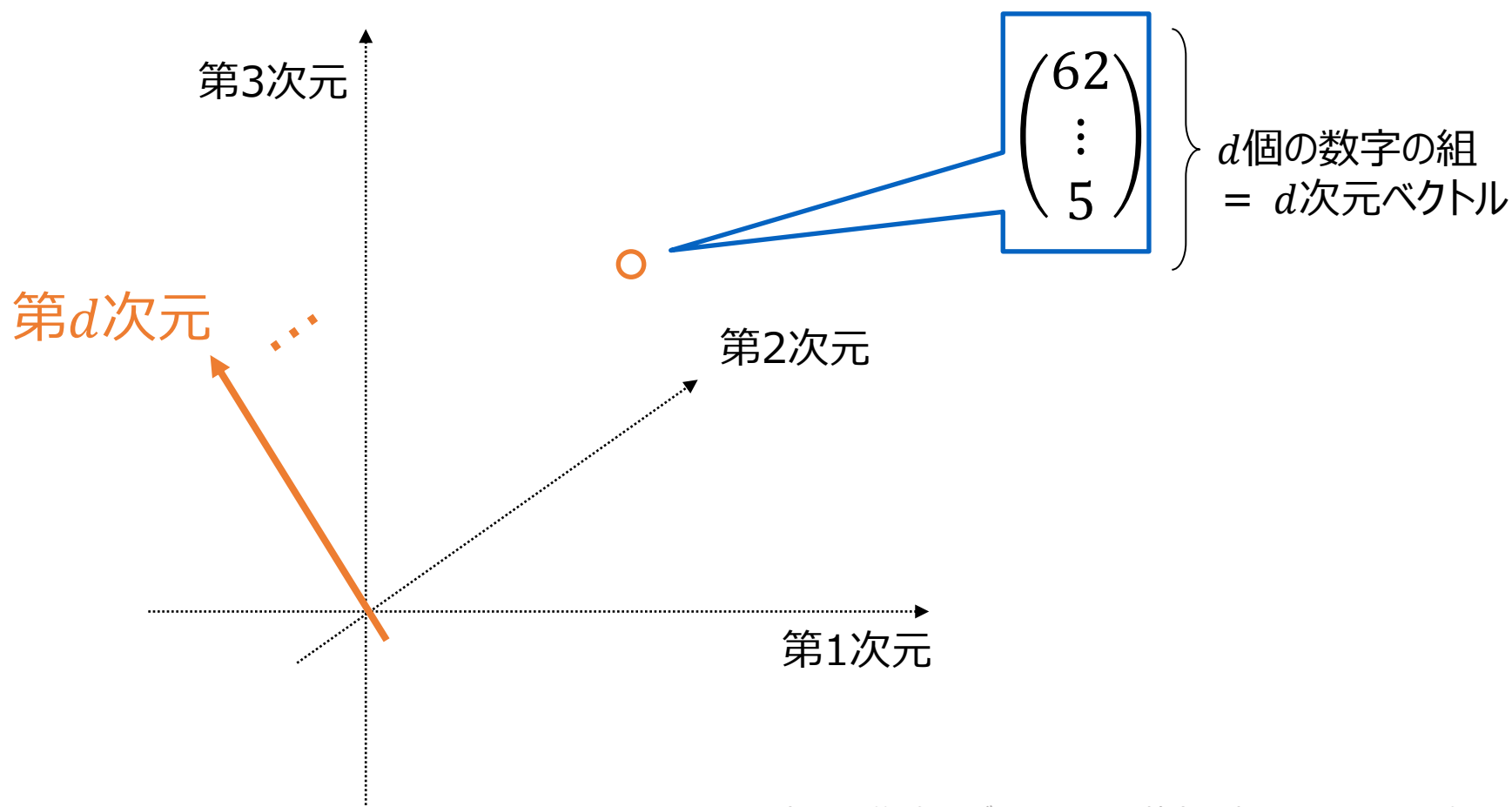
ベクトルと座標系： 「イイカゲン表現」を許してくれるば、やはり「1データ=1点」

- 図中での点の位置は「イイカゲン」ではあるが...



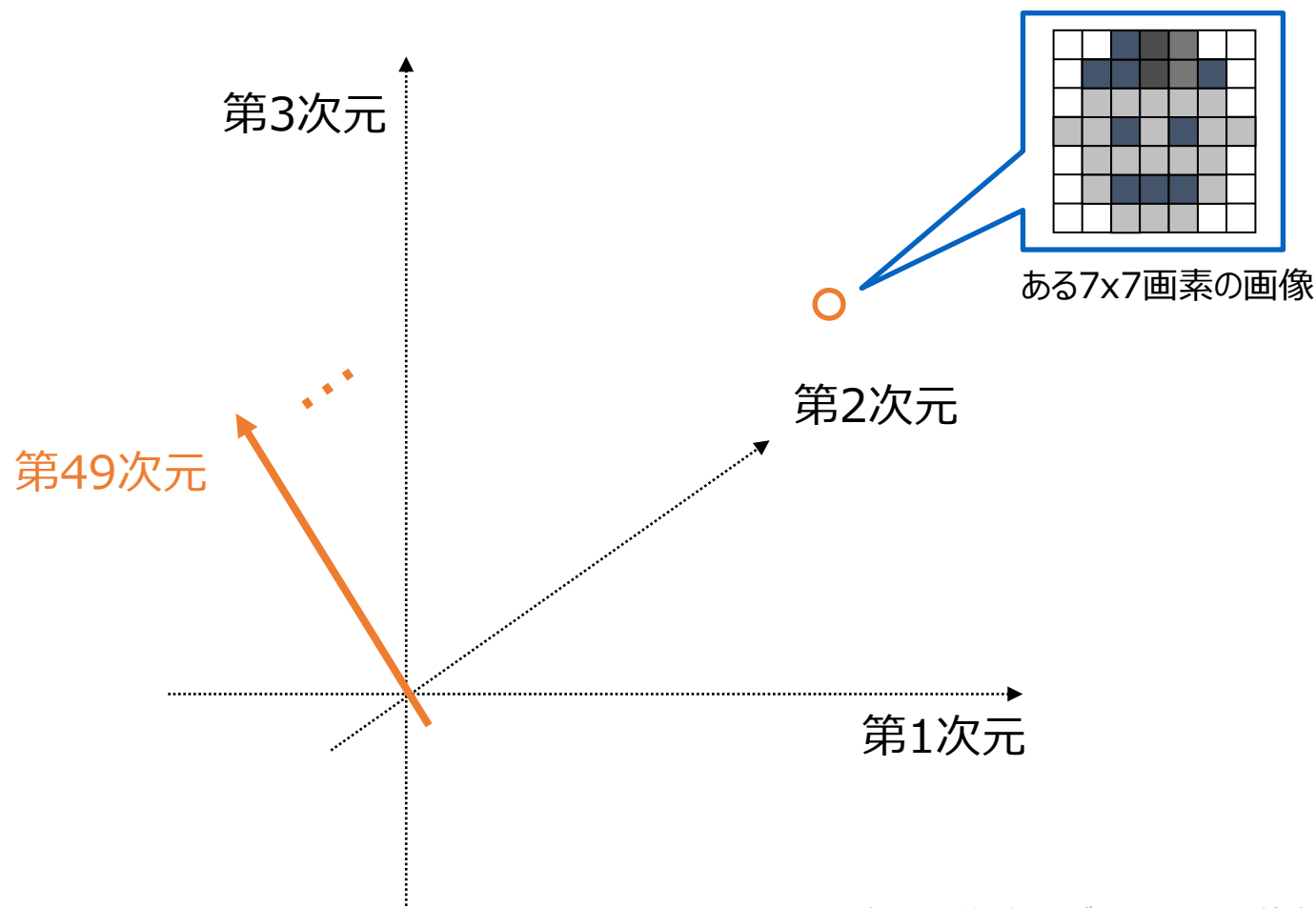
ベクトルと座標系： 4次元を許してくれたのなら、もっと多次元も許して

- 「気に入らん！」と思うかもしれませんが、そこを何とか....

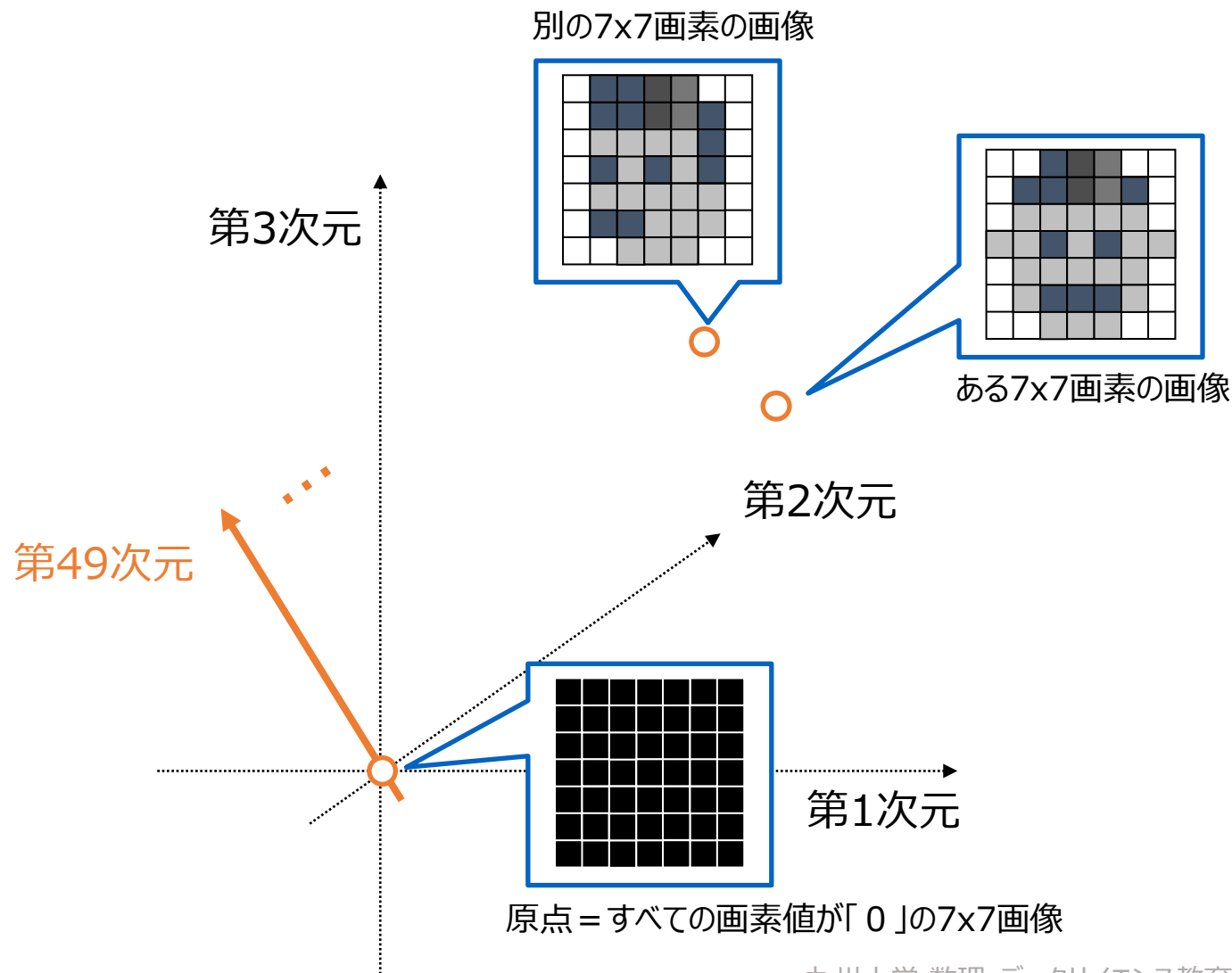


ベクトルと座標系： この勢いで、各画像データも1点にしておもう

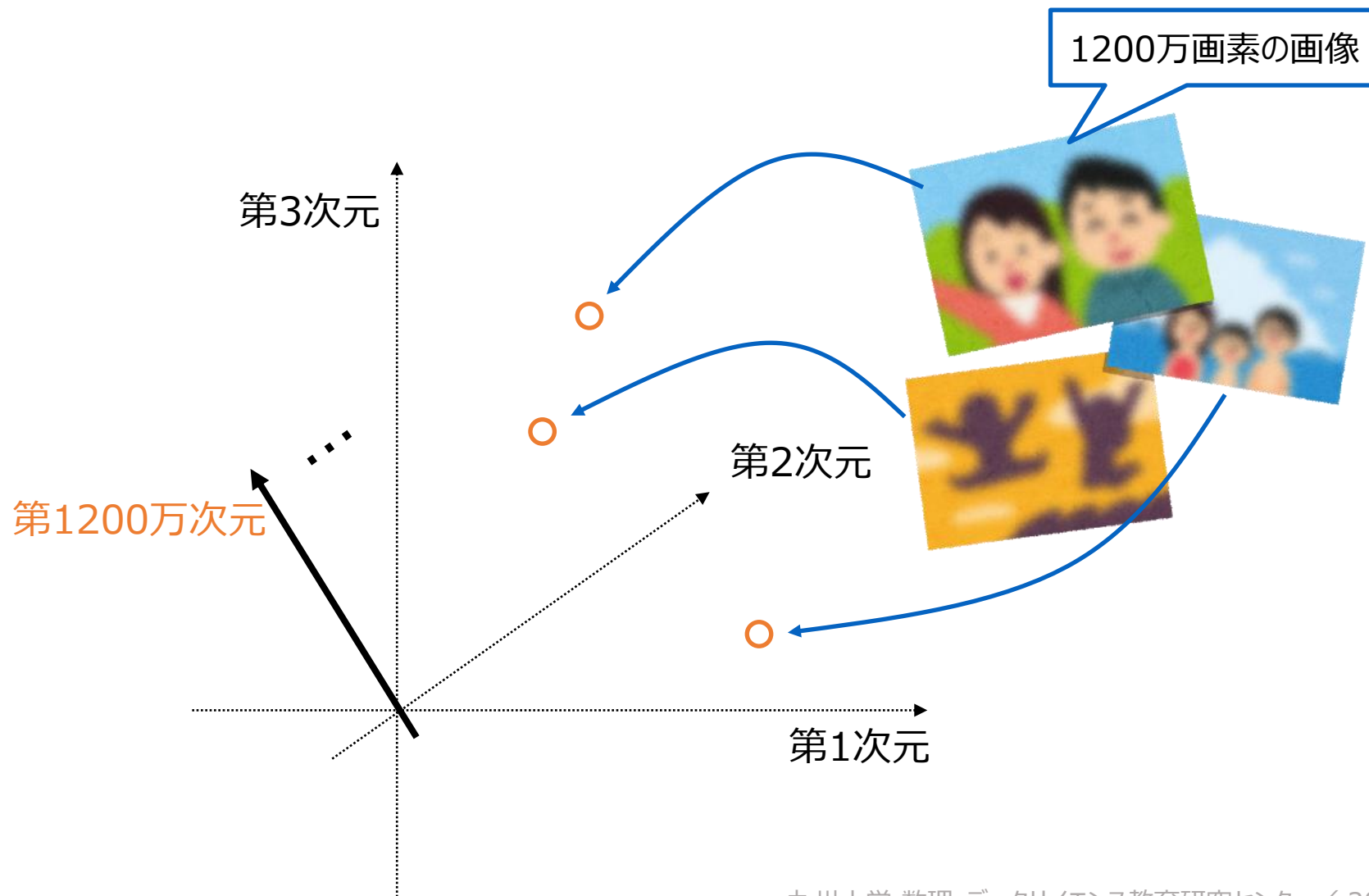
- $d = 49$ としたら、7x7画像が1点に！



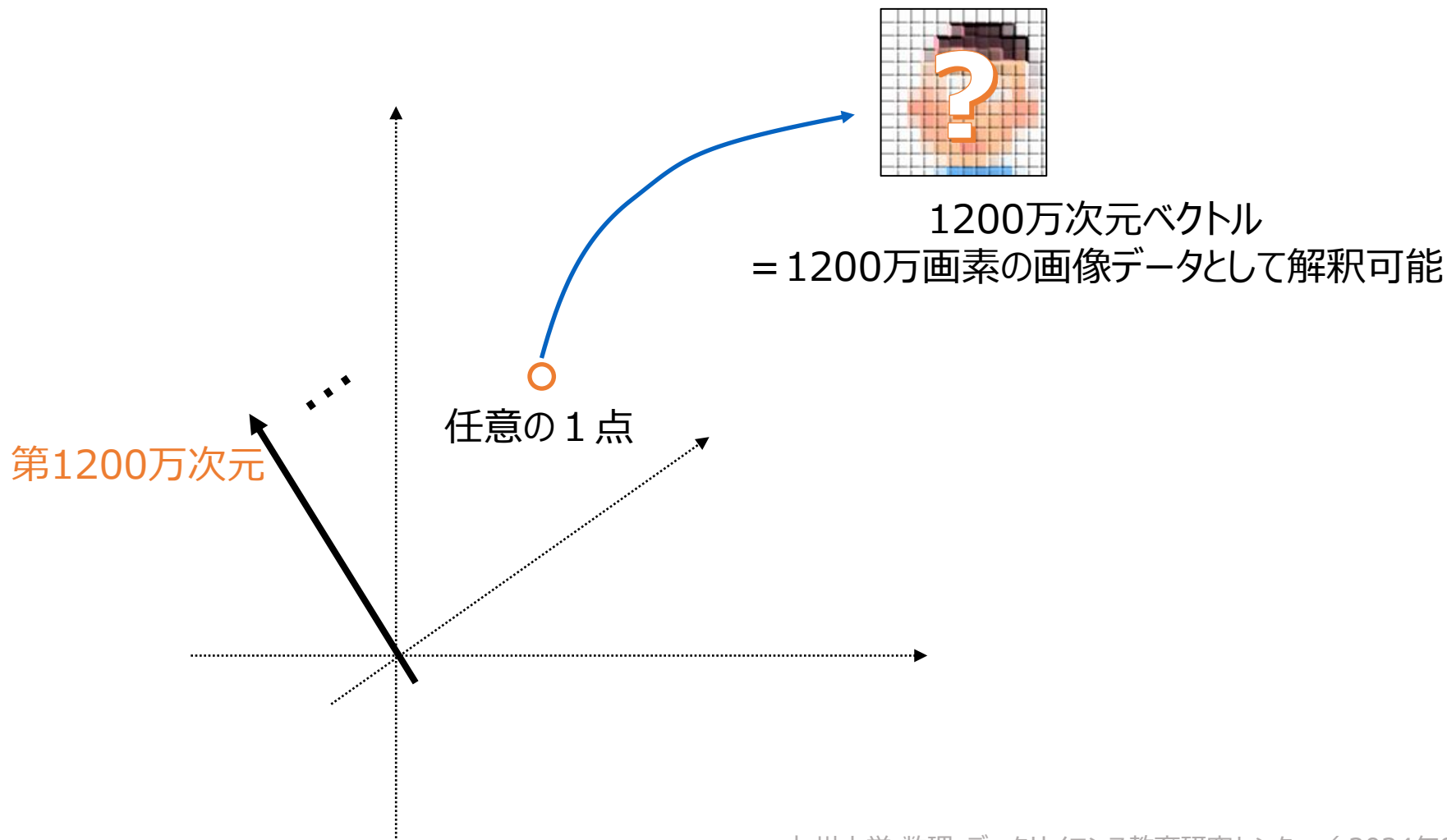
ベクトルと座標系： この勢いで、各画像データも1点にしてしまおう



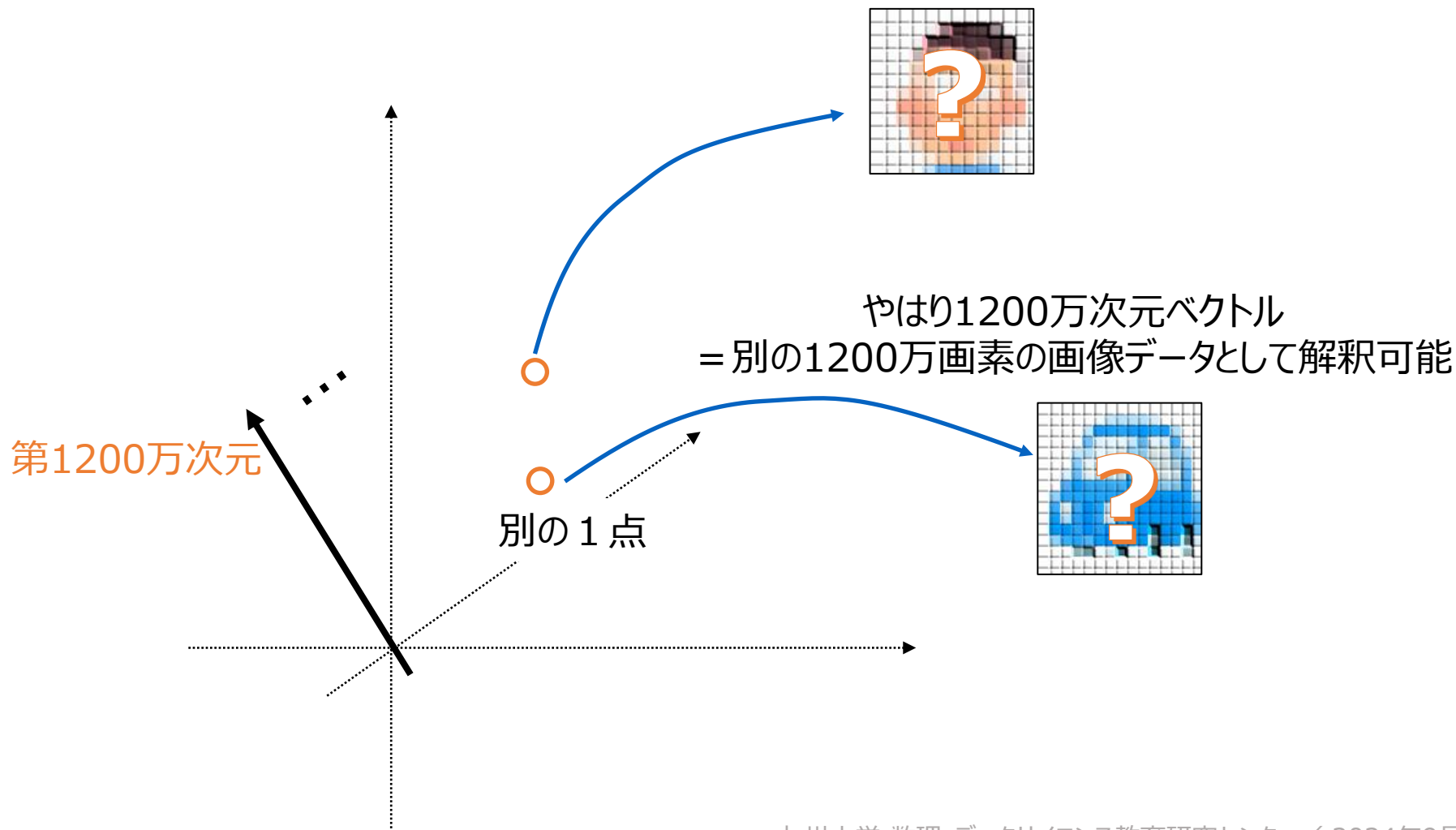
ベクトルと座標系：
もう慣れましたか？ 任意の画像についてもOKですよ！



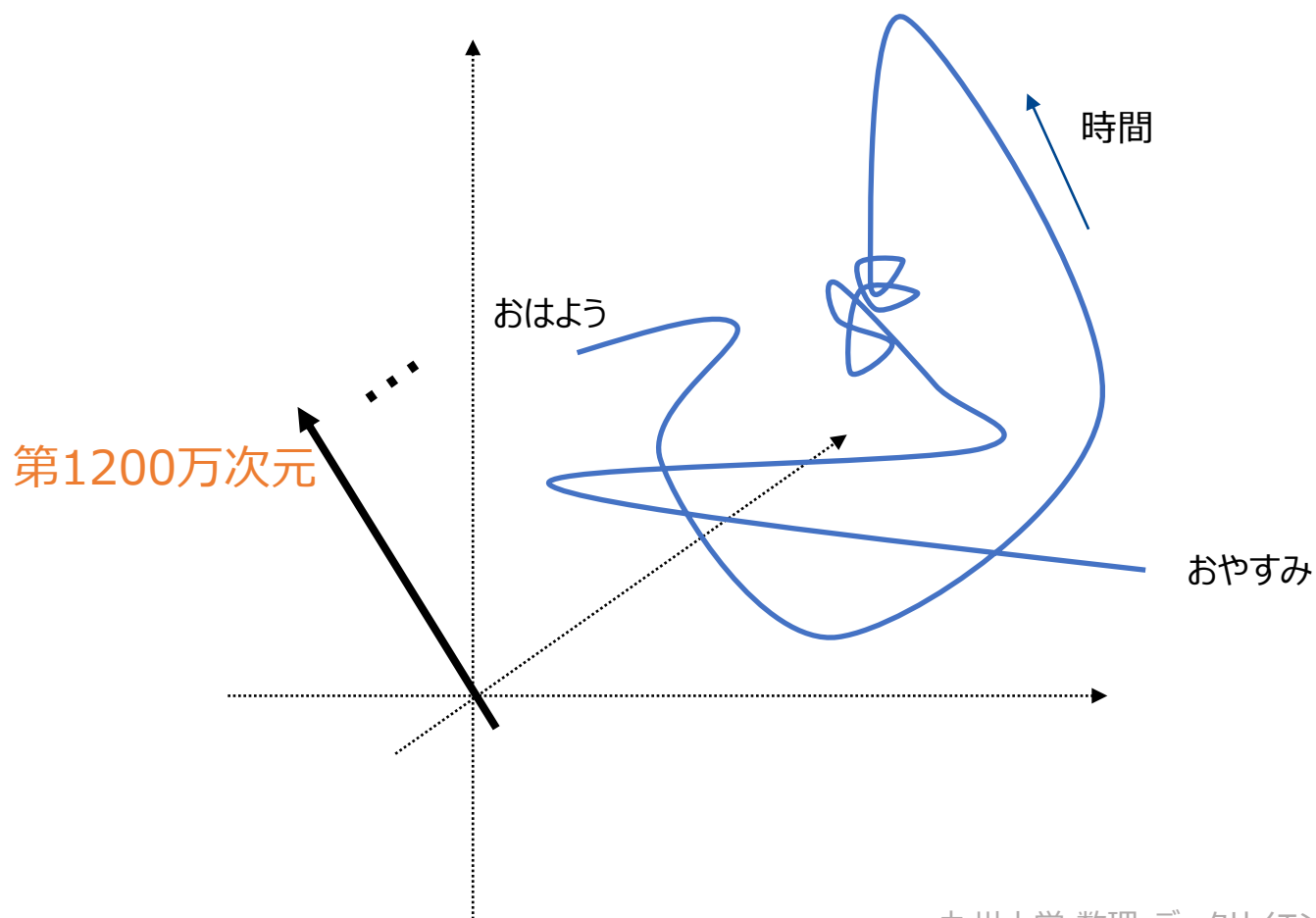
ベクトルと座標系： 逆に考えれば、任意の1点は何らかのデータに対応



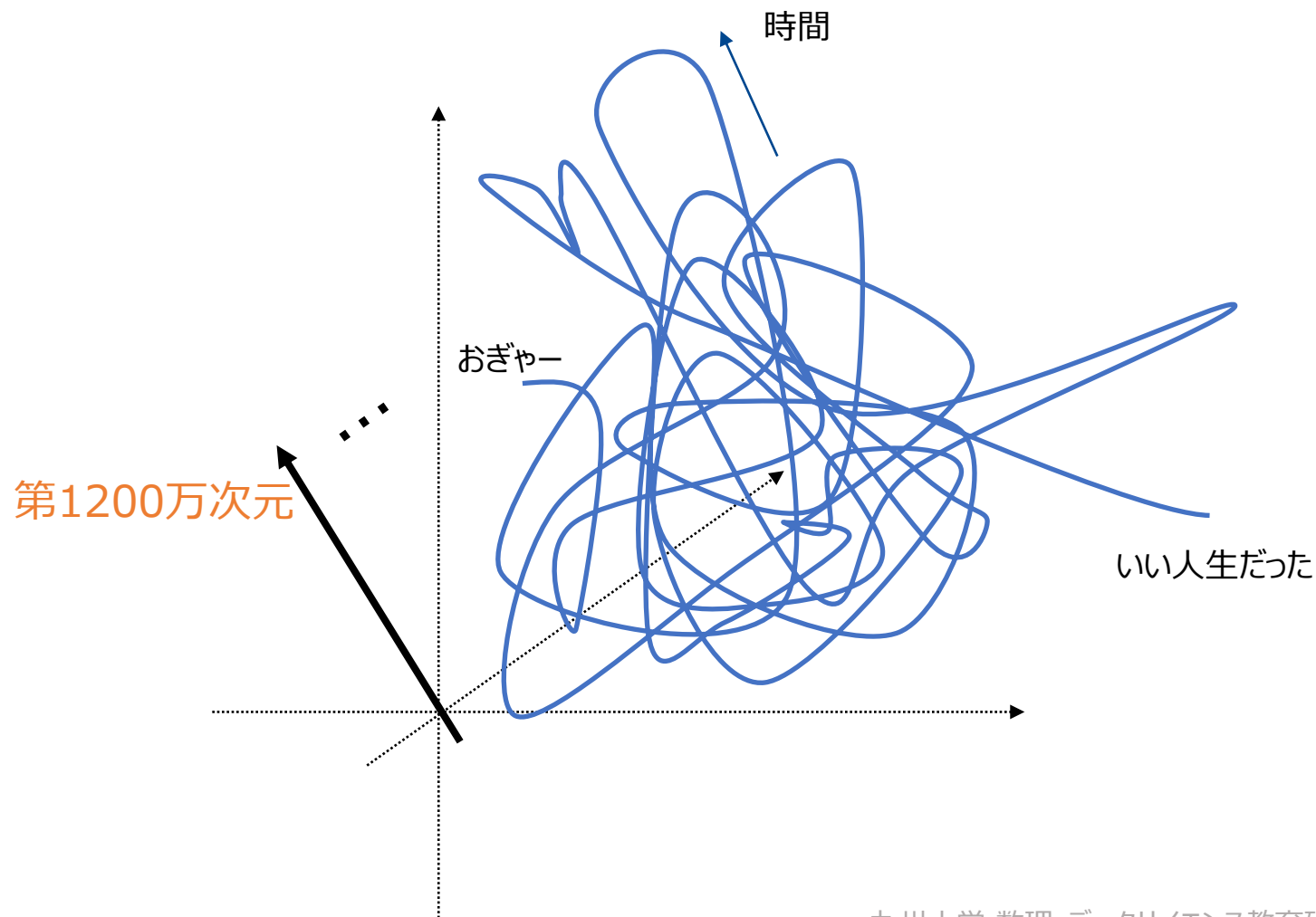
ベクトルと座標系： 逆に考えれば、任意の1点は何らかのデータに対応



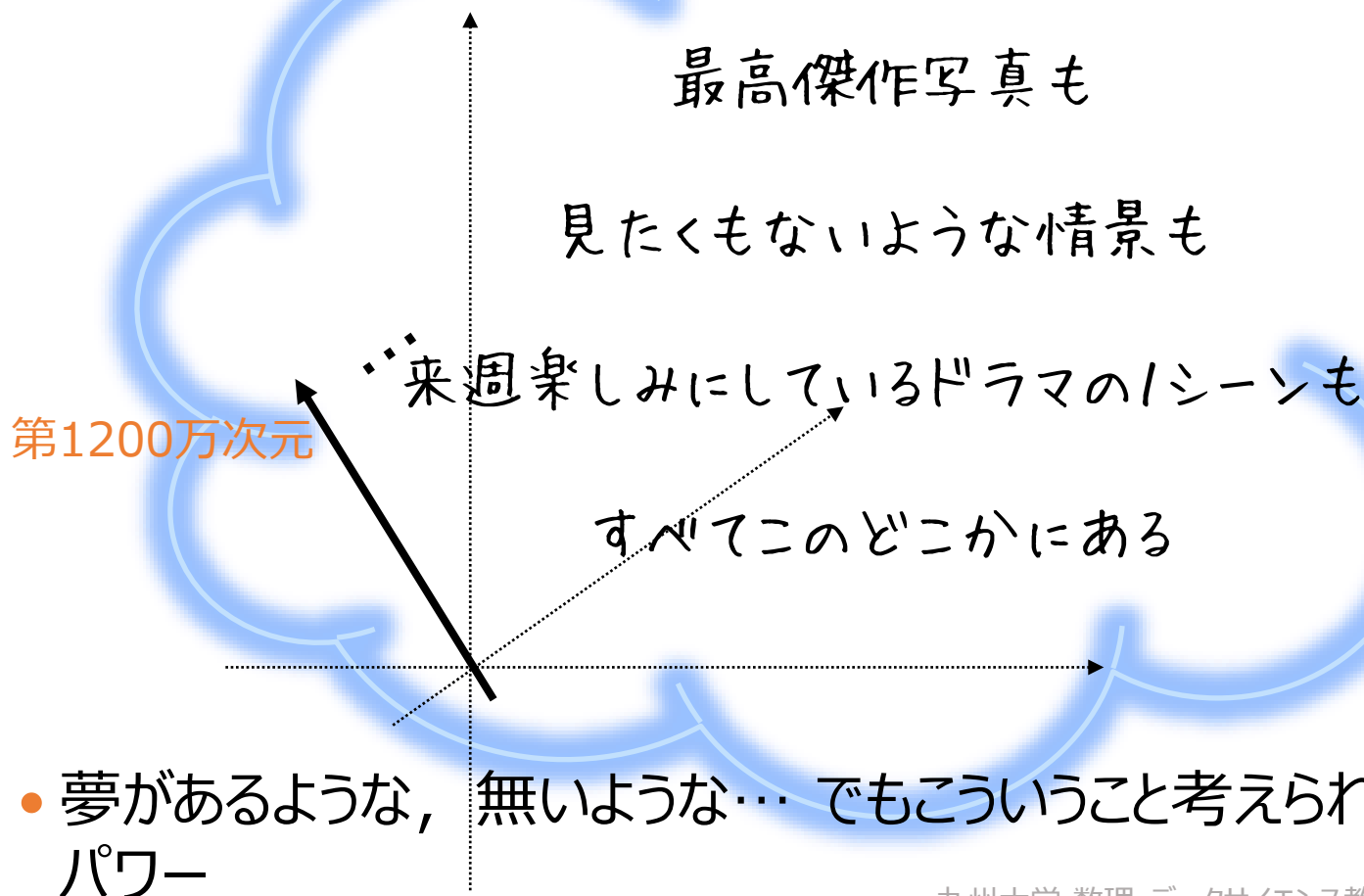
余談：皆さんが1日で目にするもの (皆さんの視覚が1200万画素だとしたら)



余談：皆さんが一生で目にするもの



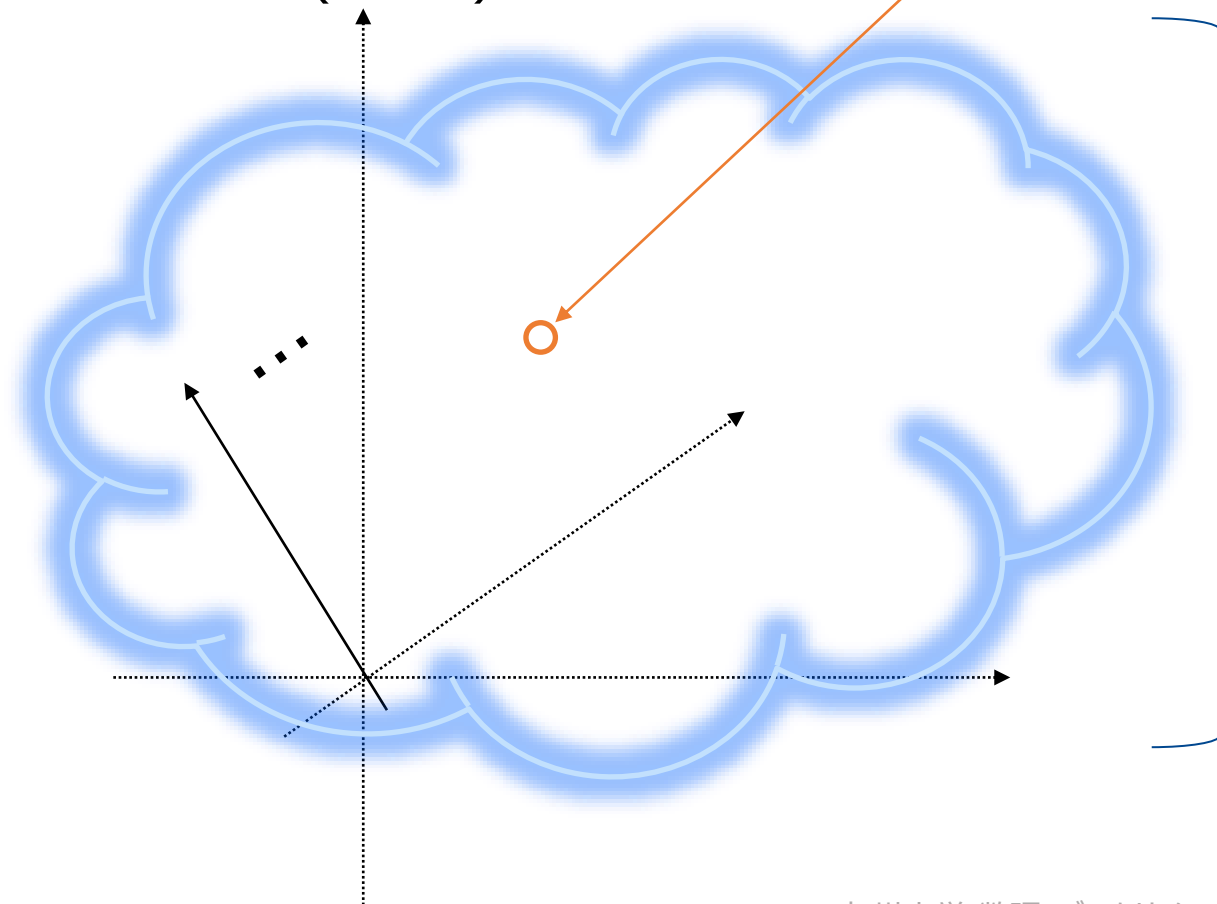
余談：要するにすべての画像(1200万画素)が この↓中に「すでに」ある！



ベクトルと座標系： この座標系の全体を「 $(d$ 次元)ベクトル空間」と呼ぶ

- (数学で出てくる)空間とは集合のこと
- ベクトル空間 = (全ての)ベクトルの集合

集合の1要素



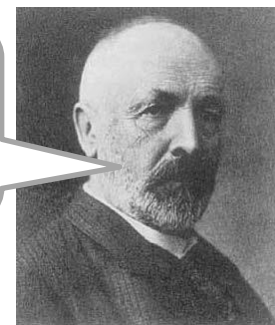
この座標系は全ての
 d 次元ベクトルが
ぎっしりつまった集合

↓
要するに(意味があ
るデータかどうかは別
として)「考え得る
全データ」の集合

余談：数学で出てくる「(ナントカ)空間」とは 何らかの「集合」のことである

- ベクトル空間 = ベクトルの集合
- 関数空間 = 関数の集合
- 線形空間 = a と b がメンバとして入っているなら $a+b$ や ca (c は定数)もメンバとして入っている集合
- 距離空間 = 要素間の距離が定義された集合
- 内積空間 = 要素間の内積が計算できるものの集合
- 位相空間 = 位相が定義できるものの集合
- バナッハ空間
- ヒルベルト空間
- などなど... (あなたも作ってみては?)

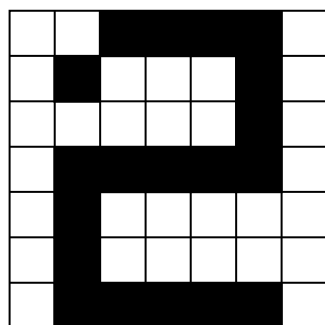
数学の本質は
その自由さにある
The essence of mathematics
is its freedom.



G. Cantor (1845-1918)

割と大事な余談： データの表現方法は一意ではない！

- 解析に適したようにデータを表現することは、重要なポイント！
- 例：



7×7画像



各行の
黒画素数を
カウント


$$\begin{pmatrix} 4 \\ 2 \\ 1 \\ 1 \\ 5 \\ 1 \\ 1 \end{pmatrix}$$

7次元ベクトル

「あなた」を表現するのは、どんなベクトルが適切？
性格だったら？ 健康状態だったら？ 知識だったら？



いよいよ数式がちょっと出ます

1つのデータ

=1つのベクトル
= d 個の数字の組

\mathbf{x} =

太字

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_d \end{pmatrix}$$

細字

d 次元ベクトル

再注：横に並べたほうが
楽そうなのに、なんで縦に
並べてる？ → 気にしない

参考：ちなみに T をつけると横に寝ます

$$\boldsymbol{x}^T = (x_1, x_2, x_3, \dots, x_d)$$

- T ：「転置」記号と呼ばれます(Transpose)
 - 工学や情報学ではよく出てきます



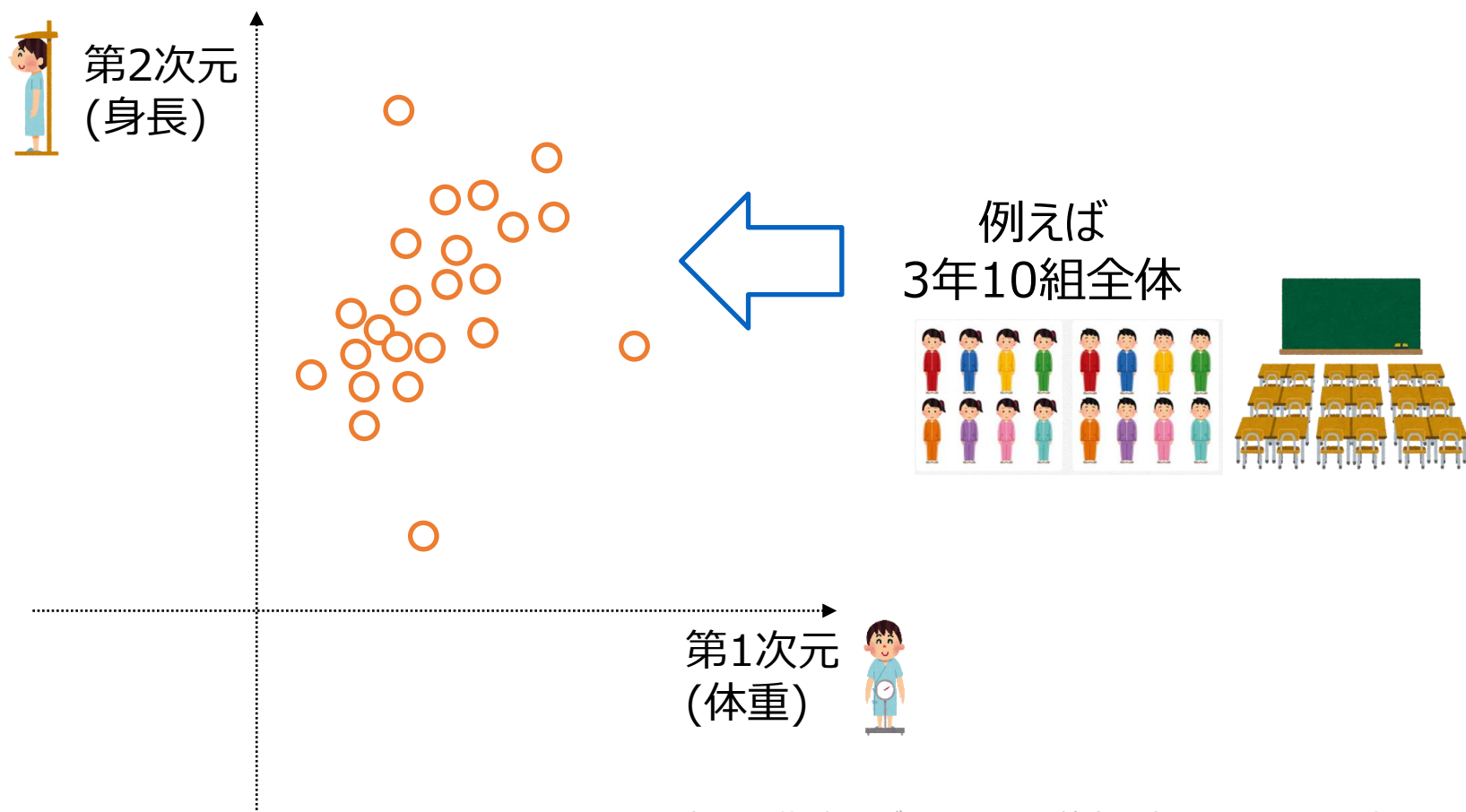
- この講義のレベルでは，寝てても立ってても，特に気にしなくてOK！
 - そのうちこれが重要になる日も来るかも…

データの集合

多くのデータを多くの点として見る

データの集合 = ベクトルの集合： 図で書くと…

- 「生徒の(体重, 身長)データのクラス全体での集合」を図で書くと



データの集合 = ベクトルの集合： 記号で書くと…

N 個のデータ

$x_1, x_2, x_3, \dots, x_N$



それぞれが d 次元ベクトル



Don't worry.
例が次スライドに

身長と体重なら....

N 人分のデータ



$x_1, x_2, x_3, \dots, x_N$



$\begin{pmatrix} 62 \\ 173 \end{pmatrix}$

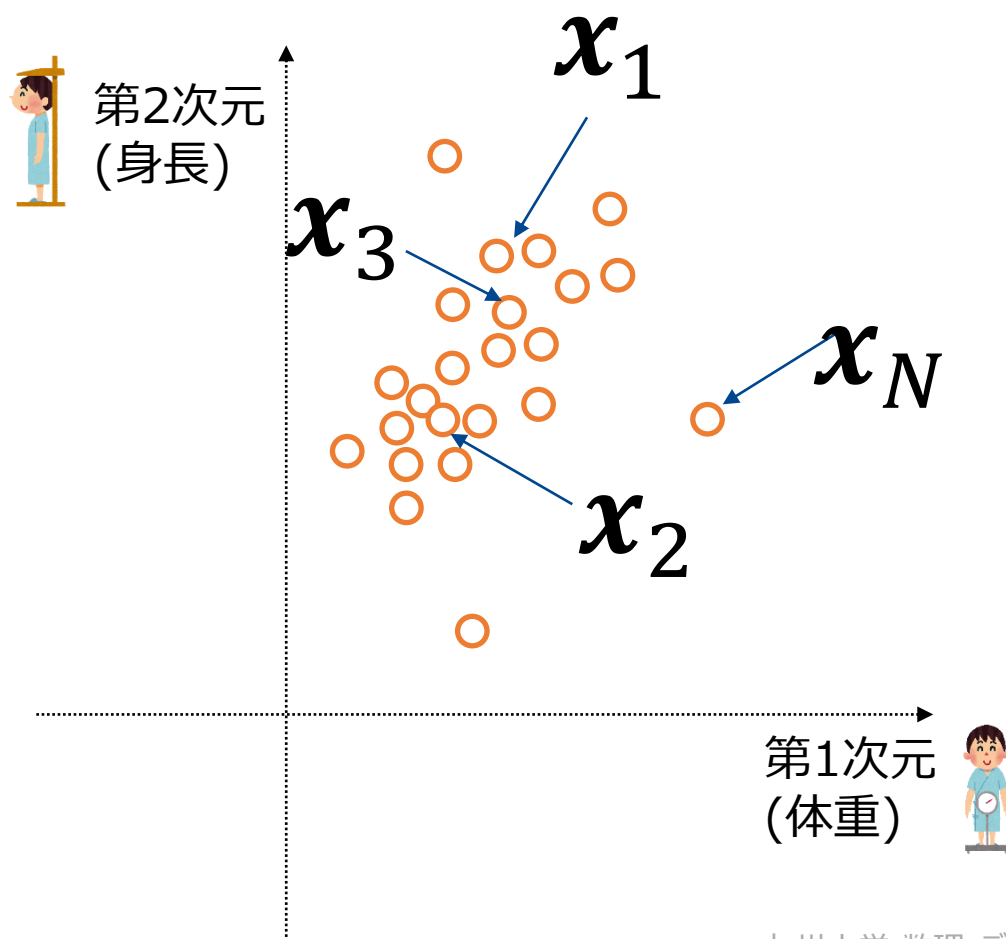
$\begin{pmatrix} 57 \\ 164 \end{pmatrix}$

$\begin{pmatrix} 65 \\ 171 \end{pmatrix}$

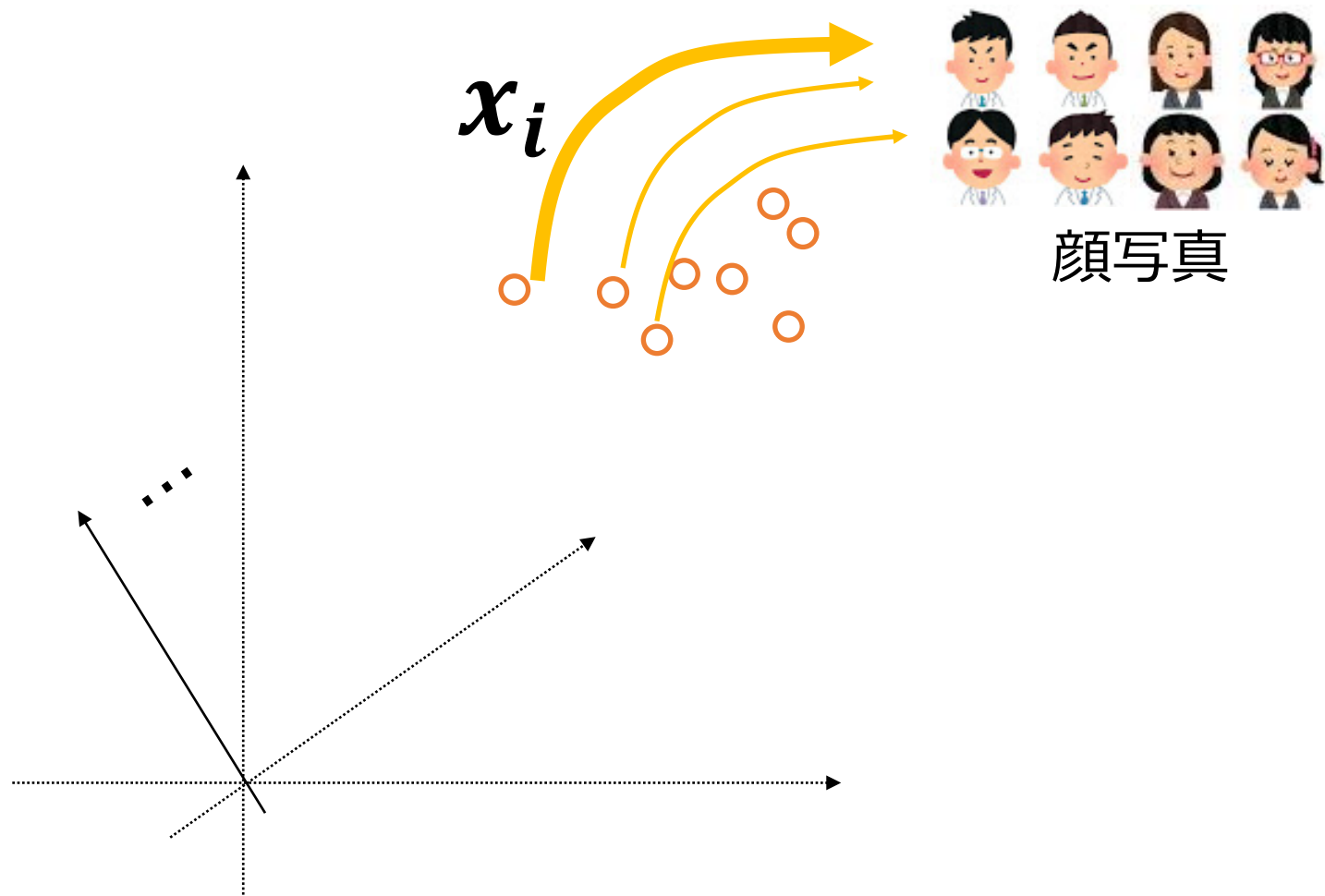
$\begin{pmatrix} 75 \\ 164 \end{pmatrix}$

それぞれが2次元ベクトル

図と記号の関係はこんな感じ







画像のような高次元ベクトルの集合でも同様



$(d\text{画素})$ 画像の空間


表にもベクトルが潜んでいる…

	 1	 2	 3	...	 N
体重	62	57	65	...	75
身長	173	164	171	...	164



1つの2次元ベクトル

もっと大きな表には高次元ベクトルが潜んでいる



	1	2	3	4	5	6	7	8
体力	0.717395	0.452726	0.715137	0.707858	0.52632	0.548061	0.853328	0.098218
知力	0.207152	0.690322	0.445059	0.428873	0.07591	0.928003	0.915683	0.723223
貯金	0.996054	0.771617	0.843382	0.017598	0.835931	0.81505	0.912822	0.347398
積極性	0.015336	0.532141	0.365665	0.295555	0.741355	0.189625	0.659545	0.228034
モテ度	0.439569	0.973285	0.30875	0.539267	0.954217	0.846784	0.220322	0.390423
器用さ	0.463235	0.895669	0.253336	0.687957	0.21026	0.241558	0.682563	0.664164
記憶力	0.200746	0.641429	0.52944	0.014236	0.064644	0.956398	0.579882	0.309734
将来性	0.680522	0.574424	0.108825	0.092599	0.888577	0.393175	0.549539	0.109326
決断力	0.265904	0.060323	0.53006	0.23987	0.710042	0.810111	0.536789	0.029784

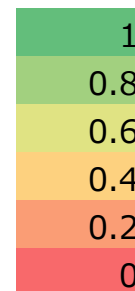


1つの9次元ベクトル

おまけ：カースケールを入れると、 大きな表もフレンドリーに



	1	2	3	4	5	6	7	8
体力	0.717395	0.452726	0.715137	0.707858	0.52632	0.548061	0.853328	0.098218
知力	0.207152	0.690322	0.445059	0.428873	0.07591	0.928003	0.915683	0.723223
貯金	0.996054	0.771617	0.843382	0.017598	0.835931	0.81505	0.912822	0.347398
積極性	0.015336	0.532141	0.365665	0.295555	0.741355	0.189625	0.659545	0.228034
モテ度	0.439569	0.973285	0.30875	0.539267	0.954217	0.846784	0.220322	0.390423
器用さ	0.463235	0.895669	0.253336	0.687957	0.21026	0.241558	0.682563	0.664164
記憶力	0.200746	0.641429	0.52944	0.014236	0.064644	0.956398	0.579882	0.309734
将来性	0.680522	0.574424	0.108825	0.092599	0.888577	0.393175	0.549539	0.109326
美しさ	0.265904	0.060323	0.53006	0.23987	0.710042	0.810111	0.536789	0.029784



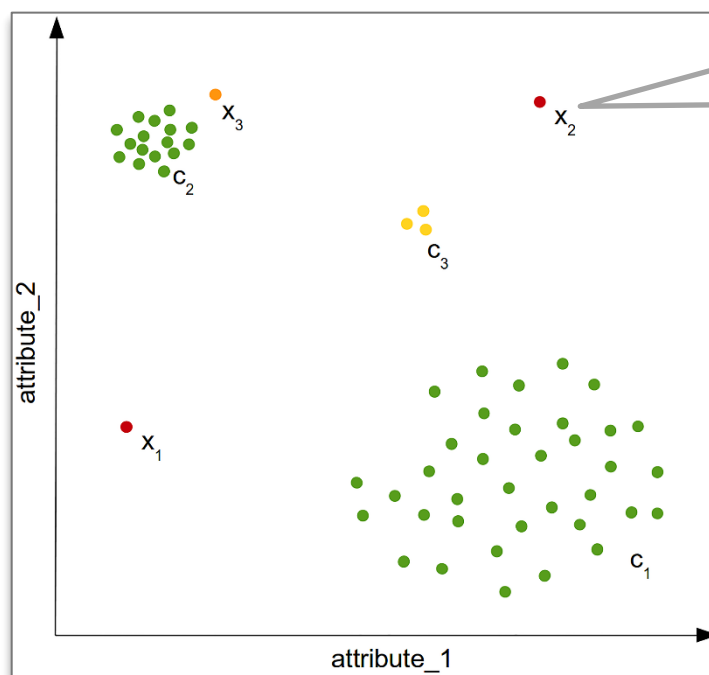
データの分布

データ集合が空間内でどう広がっているか？
→ データサイエンスのキモ！

データの分布 (1/4)

=データ集合が空間内でどう広がっているか？

- データ集合が空間内のどのような位置にどれぐらいいるか？
- この性質・傾向をきちっと数値化(定量化)すると、色々わかる！
 - 「データ解析」の基本的方法の一つ

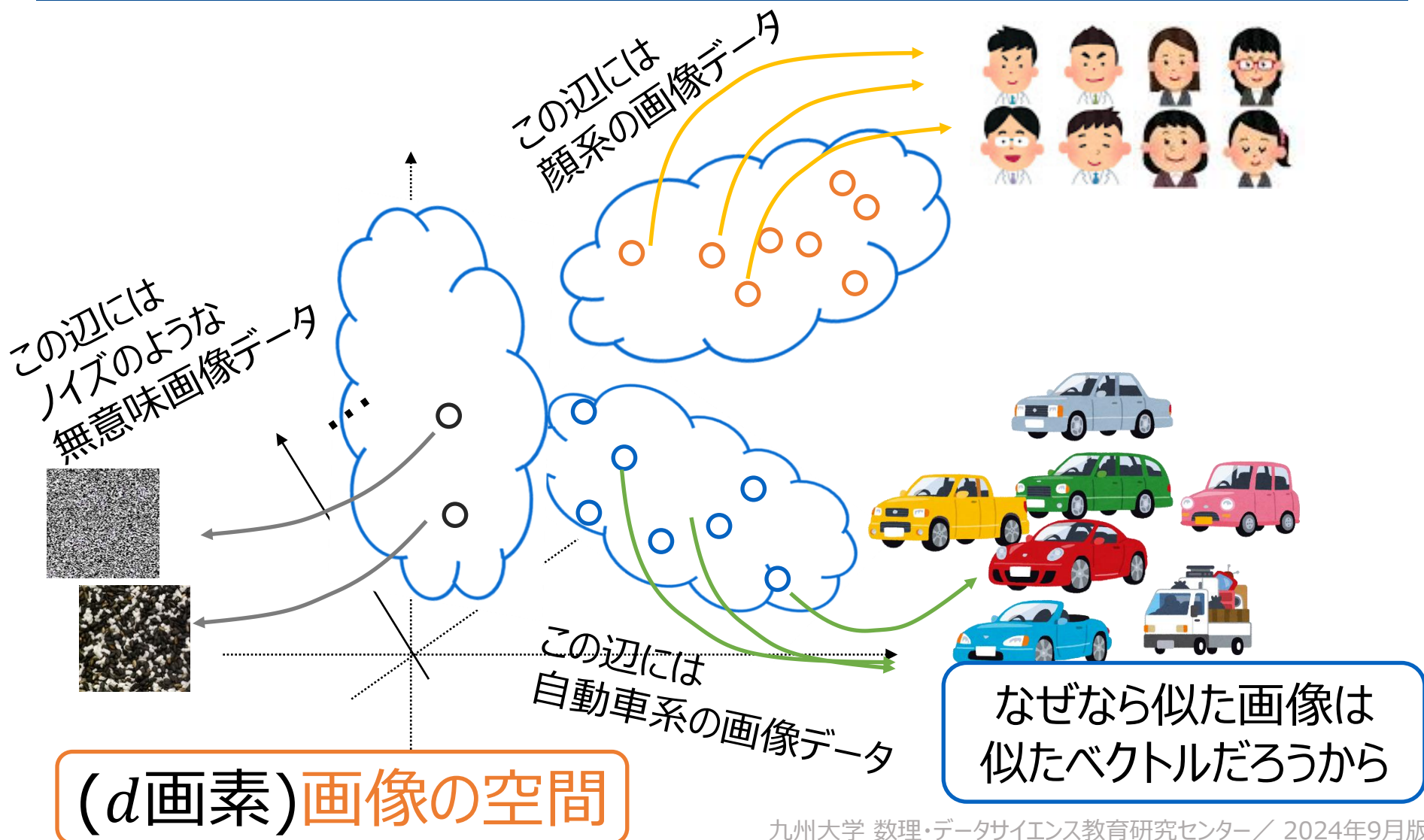


例えばこれは
「はずれデータ」
(異常かも...)

[Goldstein, Uchida, PLoSONE, 2016]

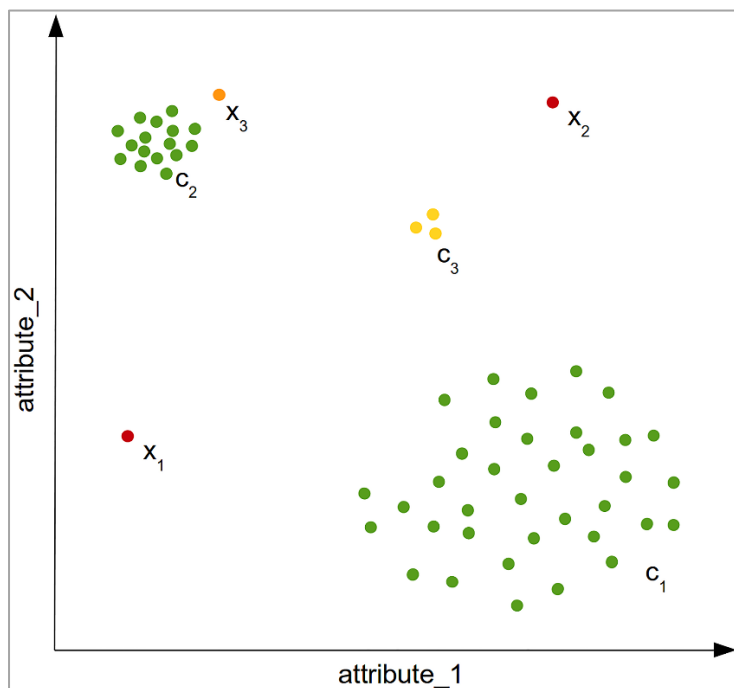
データの分布 (2/4)

似たようなデータは近くにありそう

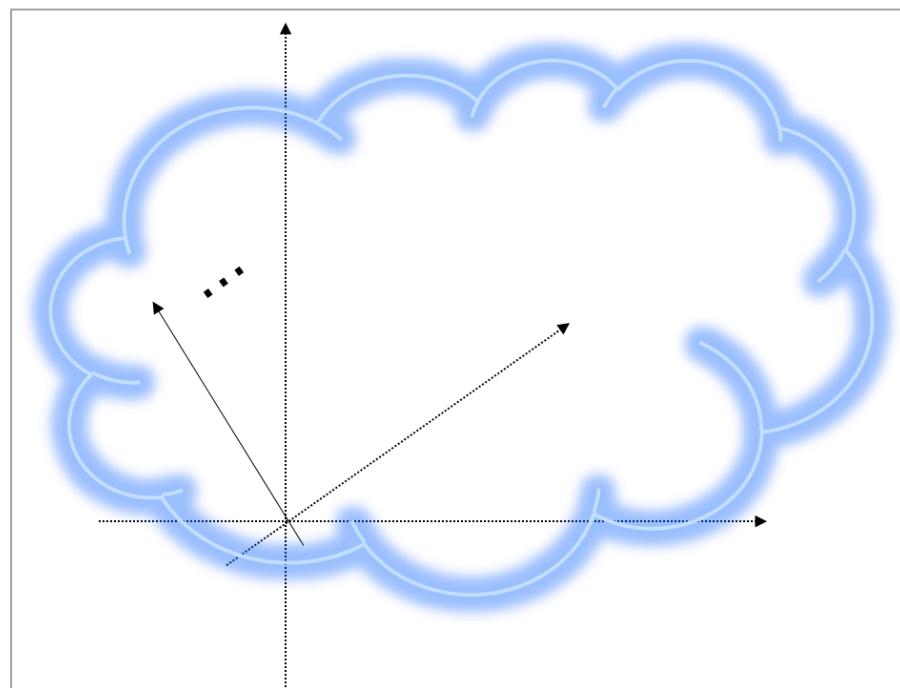


データの分布 (3/4)

高次元になると図としては可視化困難に...



2次元なら絵で描きやすく
目で見てわかる



100次元だと...!?

データの分布 (4/4)

以上より「データ分布，是非解析すべし！」

- データの分布，重要！
- 世の中には「高次元ベクトル」で表されるデータが多い
 - 「パッと絵にして理解」という方法が使えない！
- ゆえに，様々なデータ解析を行い，「データの分布状況」を何とか理解したい！
- というわけで，しばらくは「データの分布状況」理解のための方法が続きます
 - 平均，分散，クラスタリング，主成分分析...

これから出てくる方法をチラッと…

