データ分析：大別して「要約型」、「予測型」

- 次元縮約を用いない可視化
  （項目、変数が2個もしくは3個だと散布図等々
  それ以上になると次元縮約が必要
- **SVD**(singular value decomposition：特異値分解) **や PCA**（principal component analysis：主成分分析）**が要約型の代表選手**
  相関に基づく次元縮約を用いた 2D、3Dの可視化
  古典的だが基礎中の基礎
- **回帰分析**
  説明・被説明変数が所与の場合の線形代数的説明
  今回は省略．次回をご期待ください

次元縮約による分析（SVD、PCA）

・一般のデータ m 個のデータ、n 個の（観測）変数 ⇒ **mxn データ行列**

国語	社会	数学	理科	音楽	美術	体育	技家	英語
56	54	37	59	35	64	53	67	7
30	43	51	63	60	66	37	44	20
95	87	77	100	77	82	78	96	87
70	71	78	67	72	82	46	63	44
67	53	56	61	61	76	70	66	40

データ

データを俯瞰（全体的な傾向を観察できる）2Dもしくは3D 可視化

3D 表示 → 2D 表示

射影

データを「代表できる」3つの軸を使った表示

左記から「代表度」が高い2つに絞り込むんだ表示

相関行列とSVD（分散共分散行列とPCA）
まずは代表値への要約の話

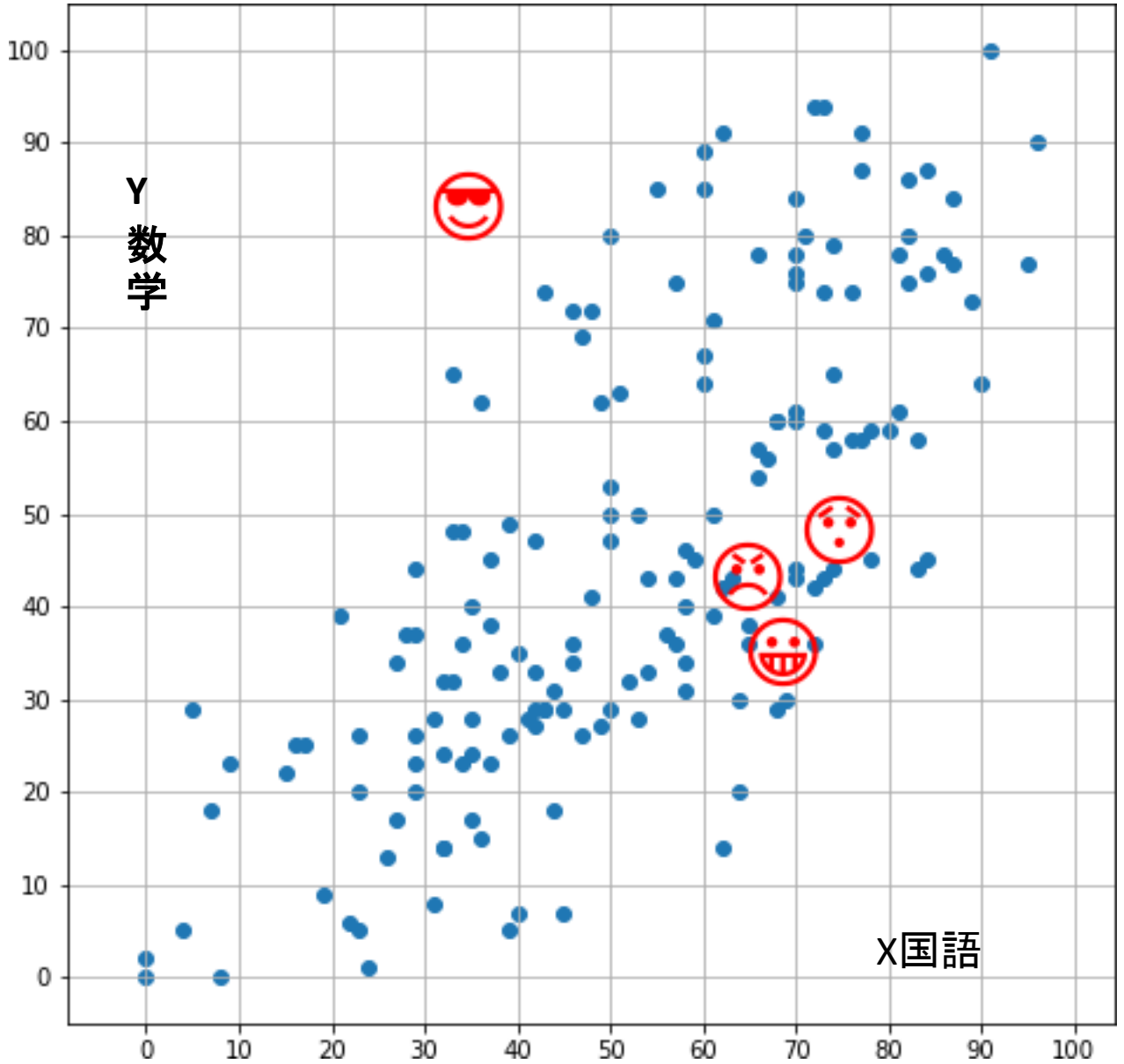
	国語X	数学Y	(X+Y)/2
😀	64	32	48
😯	70	45	58
😠	60	40	50
😎	30	80	55

＋165人のデータ

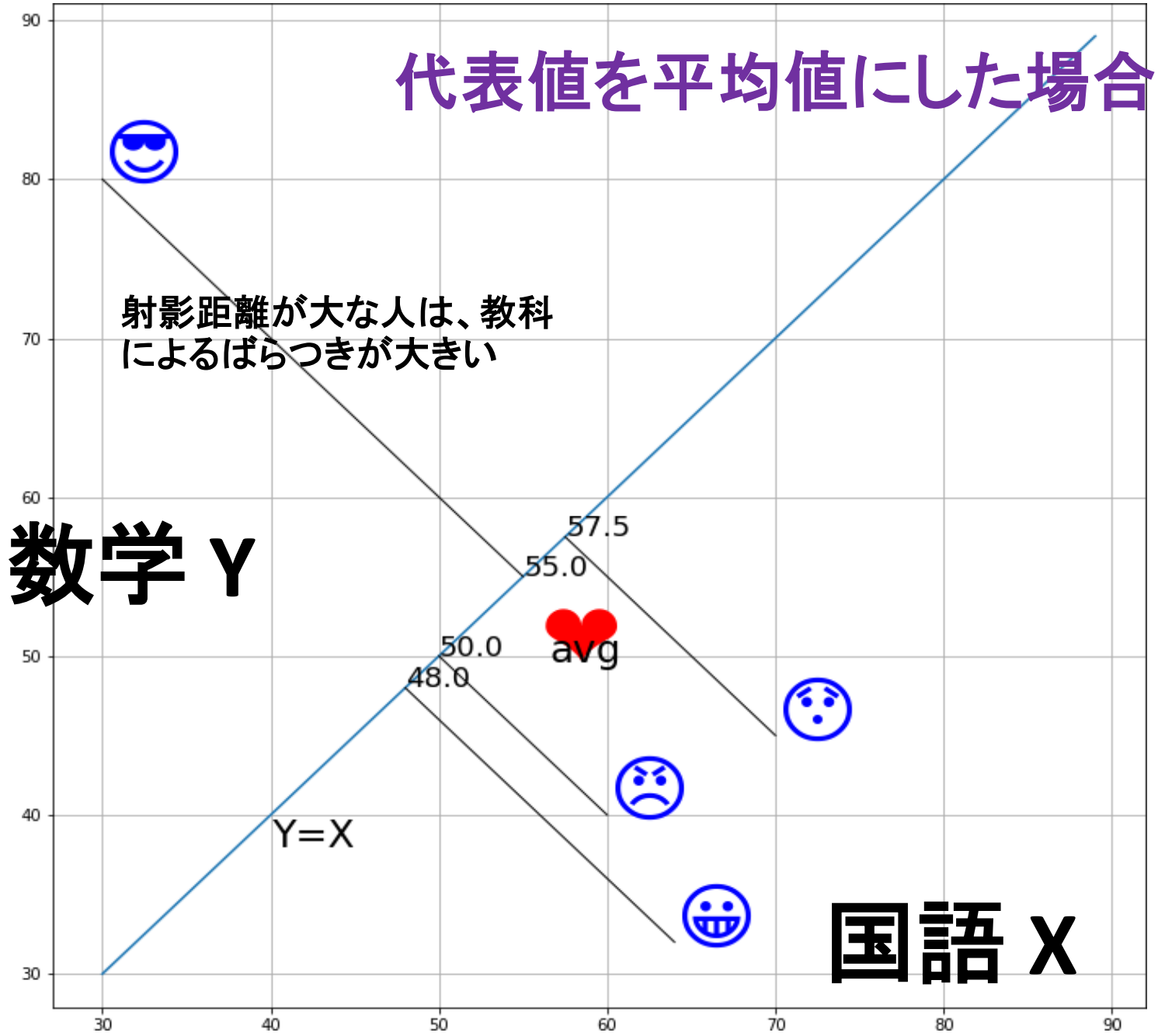
散布図で全体的な相関、および

😎 君は数学上位グループだが、
ちょっと偏りすぎなこともわかる
総合評価指標で計量できるか．．．

注目する人の顔が見える散布図



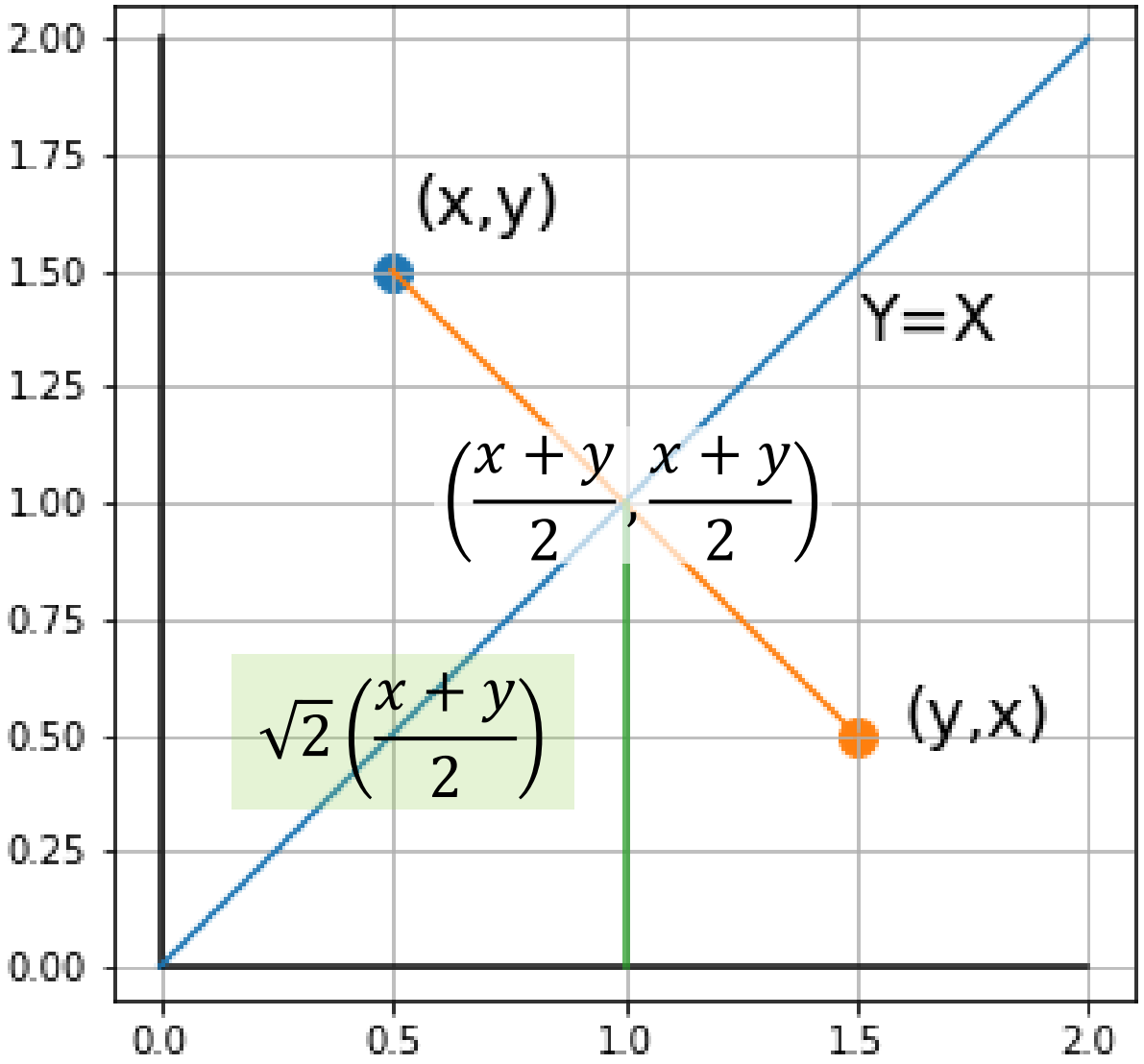
代表値を平均値にした場合



	国語 X	数学 Y	(X+Y)/2
😀	64	32	48
😯	70	45	58
😠	60	40	50
😎	30	80	55

数学が難しかったようで、
　数学に重みをつけた評価？
評価法は様々
　合理的な基準が欲しい:

各自の平均点：幾何学的に説明すると．．．



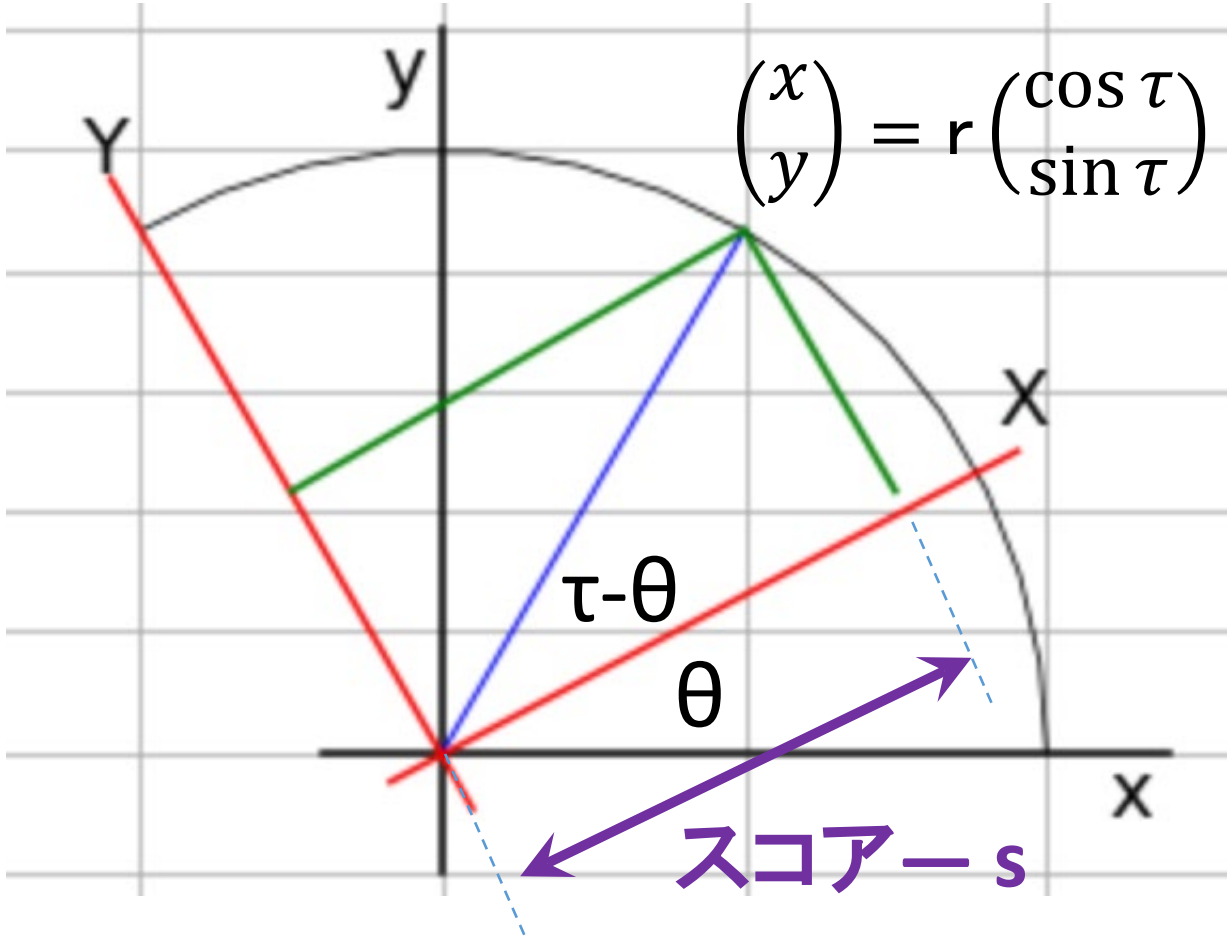
(x,y) から Y＝X 上に
垂線を下ろす

足の座標：中点

原点からの距離：
Y=X を評価軸としたときの
スコアー

合成得点（スコアー）と座標変換

傾き tan θ の軸とその直交軸に座標変換



$$\begin{pmatrix} x \\ y \end{pmatrix} = r\begin{pmatrix} \cos\tau \\ \sin\tau \end{pmatrix}$$

$$\begin{pmatrix} X \\ Y \end{pmatrix} = r\begin{pmatrix} \cos(\tau-\theta) \\ \sin(\tau-\theta) \end{pmatrix}$$

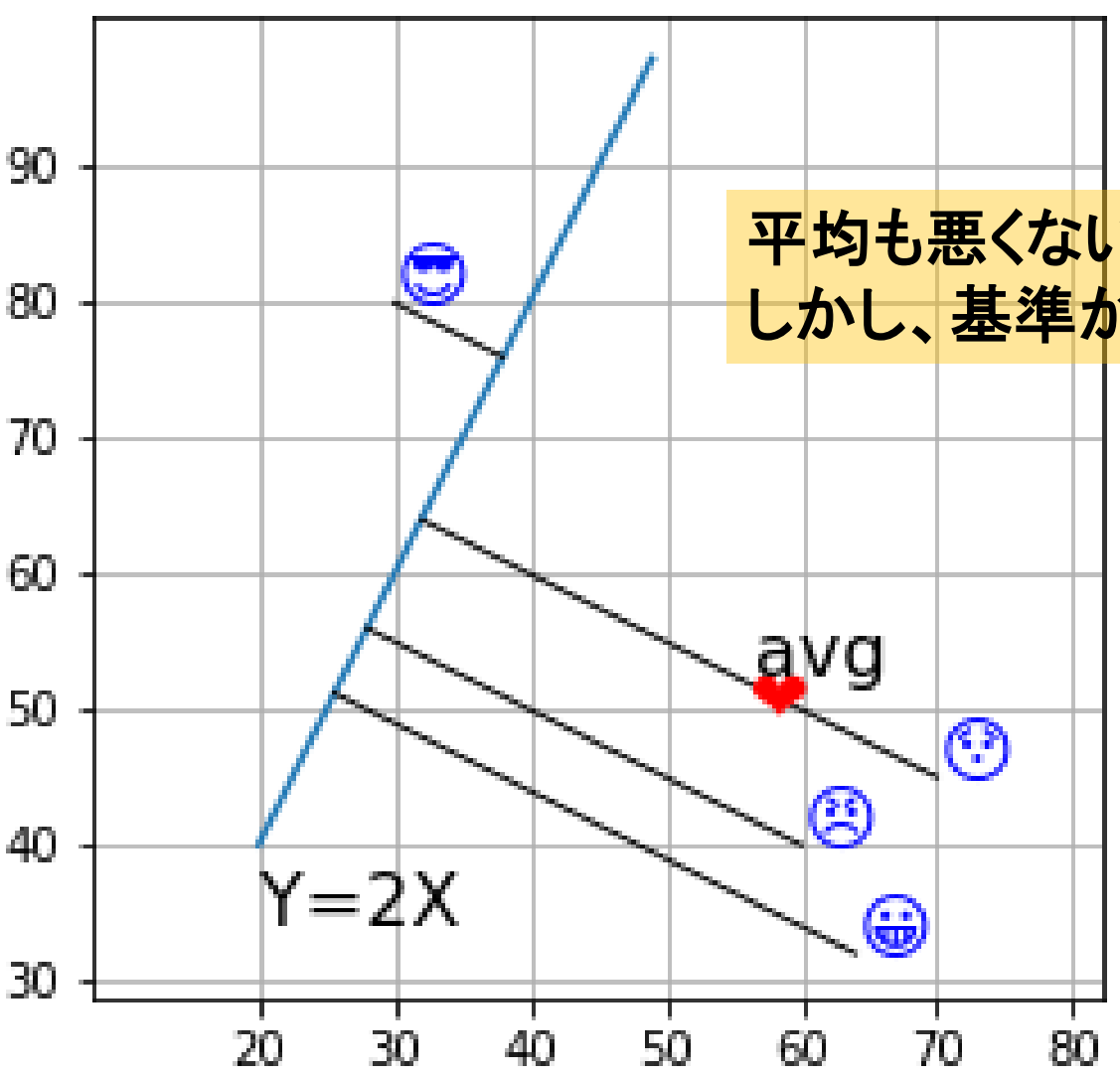
$$= r\begin{pmatrix} \cos\tau\cos(-\theta) - \sin\tau\sin(-\theta) \\ \sin\tau\cos(-\theta) + \cos\tau\sin(-\theta) \end{pmatrix}$$

$$= \begin{pmatrix} x\cos\theta + y\sin\theta \\ y\cos\theta - x\sin\theta \end{pmatrix}$$

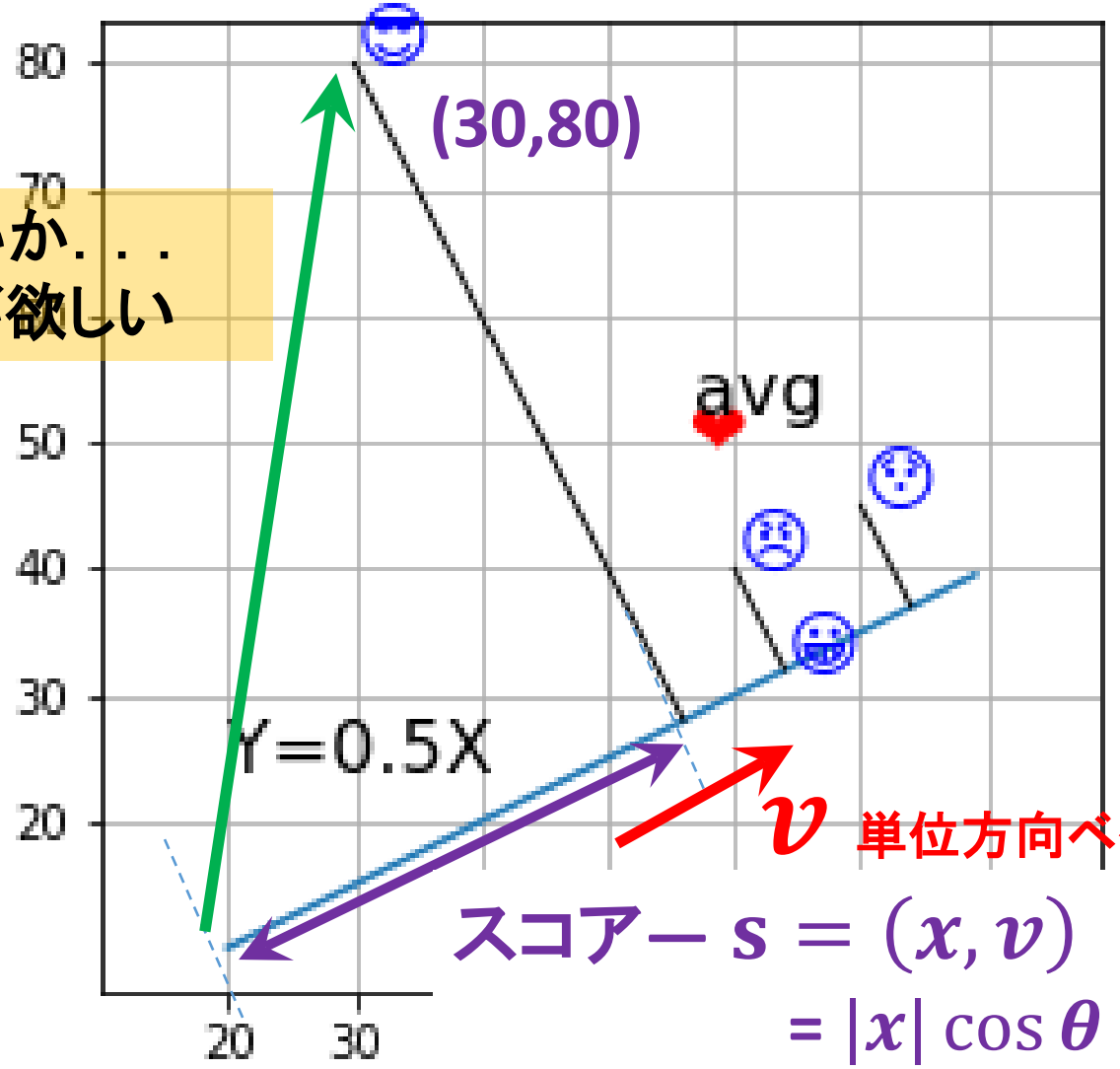
x と y の平均評価の場合

$$\theta = \frac{\pi}{4} \qquad X = \frac{\sqrt{2}}{2}(x+y)$$

平均以外は？ 傾きをいろいろ変えてみる:



平均も悪くないか．．．
しかし、基準が欲しい



v 単位方向ベクトル

スコアー $s = (x, v)$

$= |x| \cos\theta$

新変数の‘代表性’:個体の識別効果の最大化

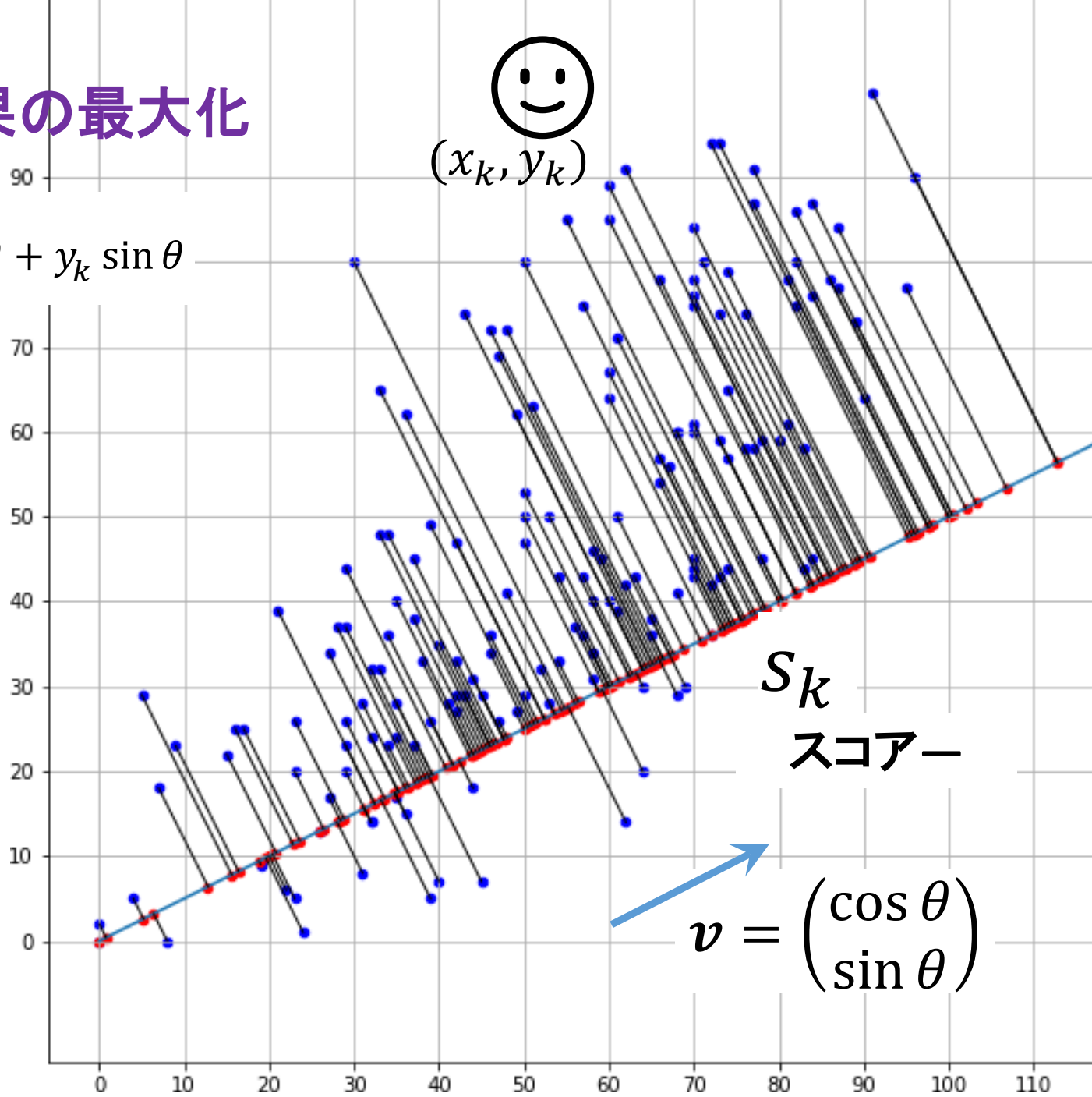
$$s_k = (x \quad y)\begin{pmatrix}\cos\theta \\ \sin\theta\end{pmatrix} = x_k\cos\theta + y_k\sin\theta$$

x	y
x_1	y_1
…	…
x_k	y_k
…	…
x_m	y_m

$$\begin{pmatrix}\cos\theta \\ \sin\theta\end{pmatrix} =$$

s
s_1
…
s_k
…
s_m

$$|Av|^2 = \sum_{k=1}^{m} {s_k}^2$$



分散に相当する量

データ行列

$$A =$$

x	y	s
x_1	y_1	s_1
…	…	…
x_k	y_k	s_k
…	…	…
x_m	y_m	s_m

$$|Av|^2 = \sum_{k=1}^{m} {s_k}^2$$

スコアーの二乗和

$$A'A = \begin{pmatrix} \boldsymbol{x}' \\ \boldsymbol{y}' \end{pmatrix} (\boldsymbol{x} \quad \boldsymbol{y}) = \begin{pmatrix} \sum_k x_k^2 & \sum_k x_k y_k \\ \sum_k x_k y_k & \sum_k y_k^2 \end{pmatrix}$$

相関行列（対称）

平均 $\mu_x = \frac{1}{m}\sum_k x_k$

分散（二乗和） $\sigma_x^2 = \frac{1}{m}\sum_k (x_k - \mu_x)^2$

$$\sigma_{xy} = \frac{1}{m}\sum_k (x_k - \mu_x)(y_k - \mu_y)$$

共分散： 項目ベクトルの内積
標準化した後は 挟み角の余弦

スコアー二乗和最大化

$$|Av|^2 = (Av)'Av = v'A'Av$$

$$A'A = \begin{pmatrix} 64 & 70 & 60 & 30 \\ 32 & 45 & 40 & 80 \end{pmatrix} \begin{pmatrix} 64 & 32 \\ 70 & 45 \\ 60 & 40 \\ 30 & 80 \end{pmatrix}$$

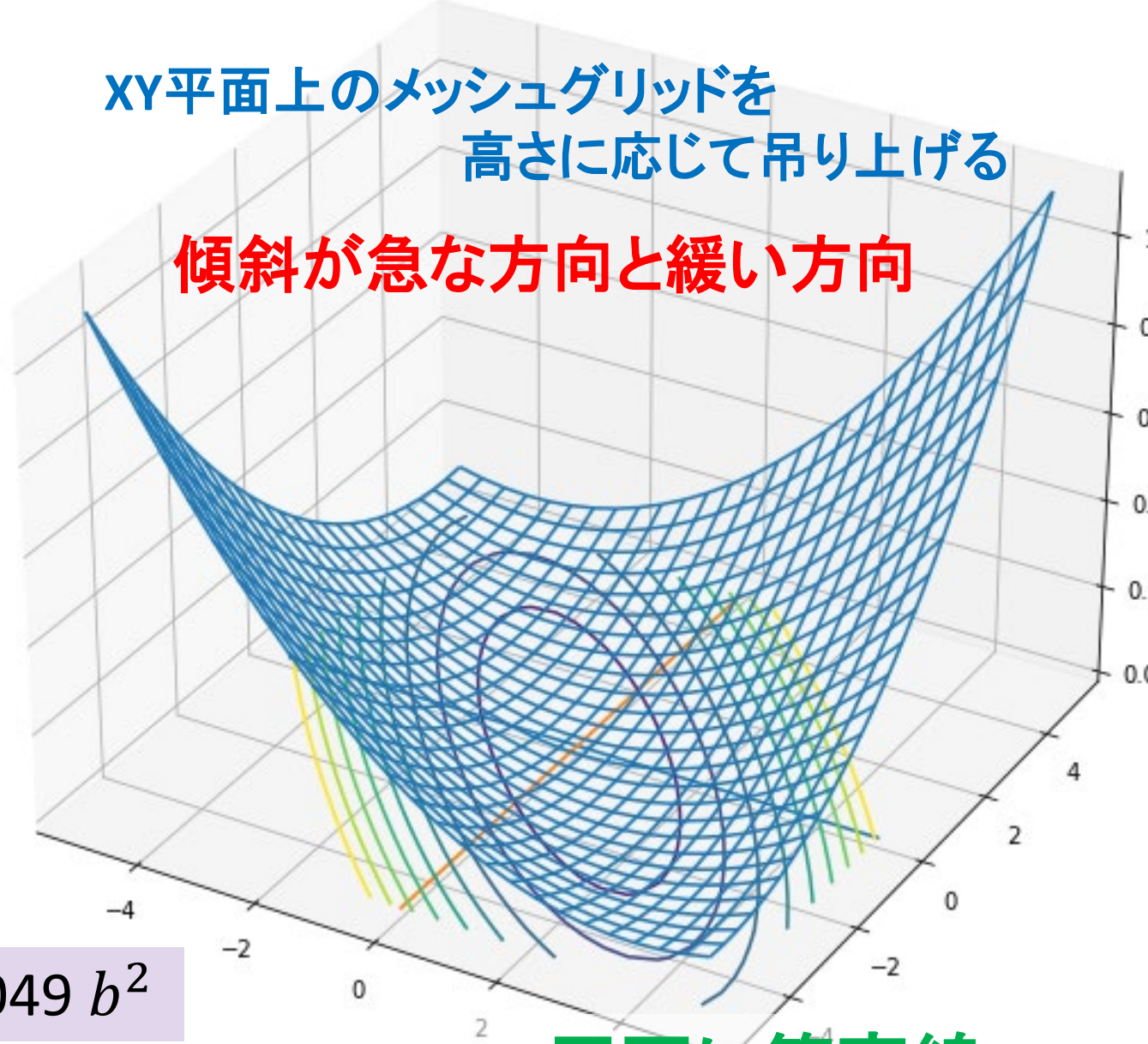
$$\begin{pmatrix} a & b \end{pmatrix} \begin{pmatrix} 13496 & 9998 \\ 9998 & 11049 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix}$$

$f(a,b) = 13496\,a^2 + 2*9998\,ab + 11049\,b^2$

相関行列の2次形式

XY平面上のメッシュグリッドを
高さに応じて吊り上げる

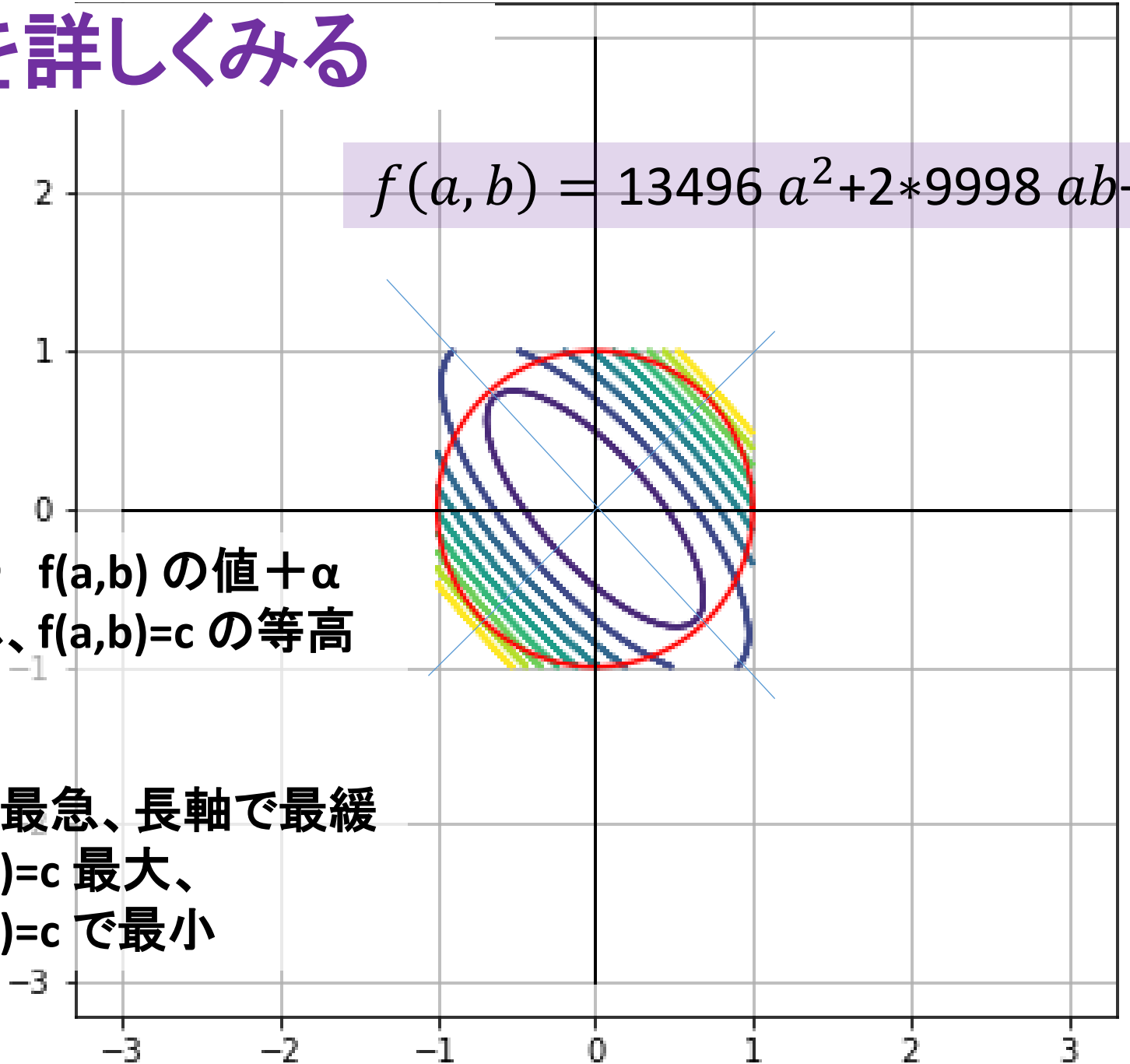
傾斜が急な方向と緩い方向



XY平面に等高線

等高線群を詳しくみる

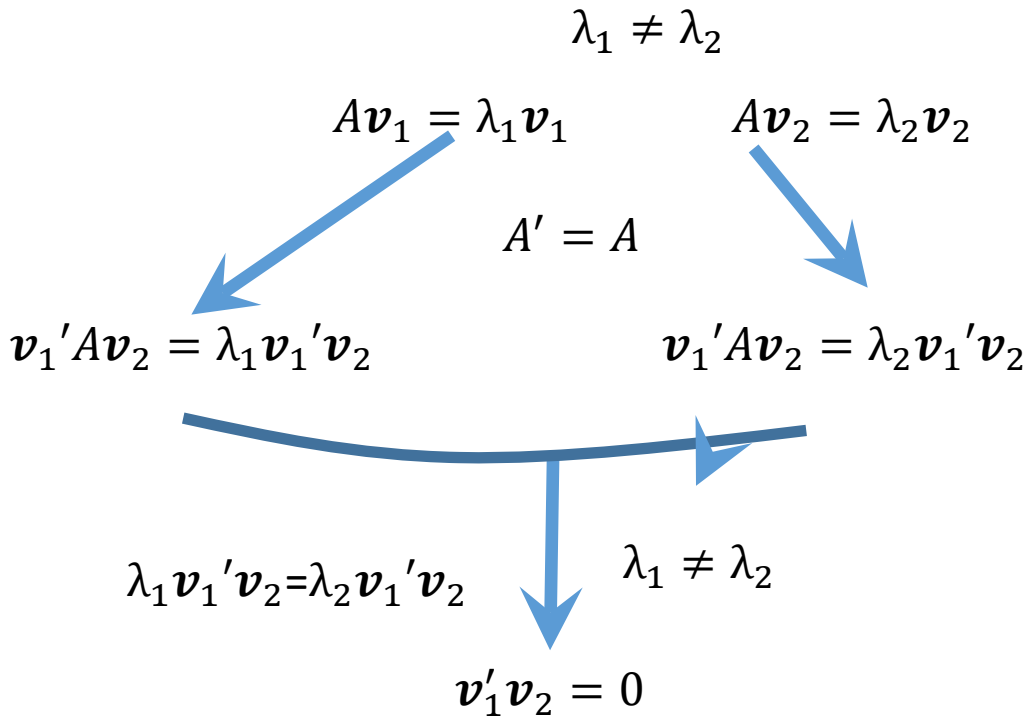
$$f(a,b) = 13496\,a^2 + 2*9998\,ab + 11049\,b^2$$



- ✓ 単位円（赤色）上の　f(a,b) の値＋α より小さな c に対し、f(a,b)=c の等高線を描いた図
- ✓ 等高線は楕円形
- ✓ 楕円の短軸方向で最急、長軸で最緩
- ✓ 円に内接する f(a,b)=c 最大、
- ✓ 円に外接する f(a,b)=c で最小

参考：$f(a,b)=c$ が楕円であること
固有ベクトルを使った座標変換

実対称行列の異なる固有値の固有ベクトルは直交



$$f(a,b)=\begin{pmatrix}a & b\end{pmatrix}A'A\begin{pmatrix}a\\ b\end{pmatrix}$$

$$A\boldsymbol{v}_1=\lambda_1\boldsymbol{v}_1 \quad A\boldsymbol{v}_2=\lambda_2\boldsymbol{v}_2$$

座標変換 $\begin{pmatrix}a\\ b\end{pmatrix}=\begin{pmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2\end{pmatrix}\begin{pmatrix}X\\ Y\end{pmatrix}$

$$\begin{pmatrix}X & Y\end{pmatrix}\begin{pmatrix}\boldsymbol{v}_1{}'\\ \boldsymbol{v}_2{}'\end{pmatrix}A'A\begin{pmatrix}\boldsymbol{v}_1 & \boldsymbol{v}_2\end{pmatrix}\begin{pmatrix}X\\ Y\end{pmatrix}$$

$$=\begin{pmatrix}X & Y\end{pmatrix}\begin{pmatrix}\lambda_1\boldsymbol{v}_1{}'\\ \lambda_2\boldsymbol{v}_2{}'\end{pmatrix}\begin{pmatrix}\lambda_1\boldsymbol{v}_1 & \lambda_2\boldsymbol{v}_2\end{pmatrix}\begin{pmatrix}X\\ Y\end{pmatrix}$$

$$=\begin{pmatrix}X & Y\end{pmatrix}\begin{pmatrix}\lambda_1{}^2 & 0\\ 0 & \lambda_2{}^2\end{pmatrix}\begin{pmatrix}X\\ Y\end{pmatrix}$$

$$=\lambda_1{}^2X^2+\lambda_2{}^2Y^2$$

= c で楕円の方程式

$$\frac{X^2}{\lambda_1{}^{-2}}+\frac{Y^2}{\lambda_2{}^{-2}}$$

短軸方向は $\lambda_1>\lambda_2$ として λ_1 の固有ベクトル　$\boldsymbol{v}_1$ 方向

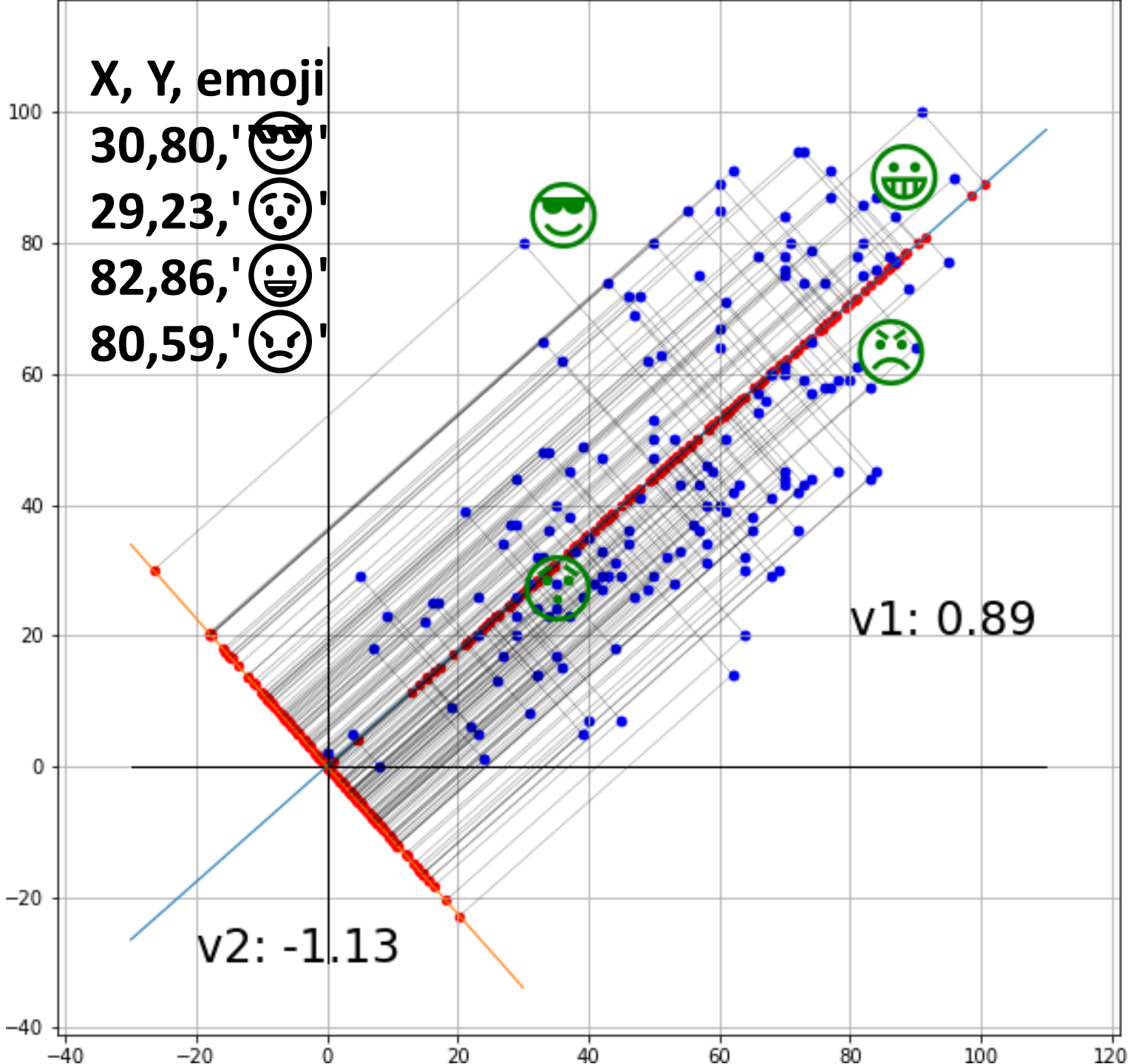
残りの（直交）軸は？

- 2Dでは分散最大化軸の
  直交軸として決まる

- 主要軸を補うという意味で
  「（直交）補空間」の生成系

この場合、ほぼ
y=x とその直交軸だが、

全体的な出来具合（v1）
＋
国語と数学の偏り具合（v2）



多変数（次元が3以上）の場合へ

- ２変数の場合は、
  そもそも**散布図を眺めれば凡その傾向は観れる**
  変数の相関（正規化すれば方向性）をみて軸を決めた.
- 多変数の場合はどうする？ 単純な可視化はもはやできない
- 考え方は同じ. **全体を俯瞰・要約できる軸**を選んでいく.
  相関行列は 2x2 から n×n に…（n は変数の数）

国語	社会	数学	理科	音楽	美術	体育	技家	英語
30	43	51	63	60	66	37	44	20
39	21	49	56	70	72	56	63	16
29	30	23	57	69	76	33	54	6
95	87	77	100	77	82	78	96	87
…	…	…	…	…	…	…	…	…

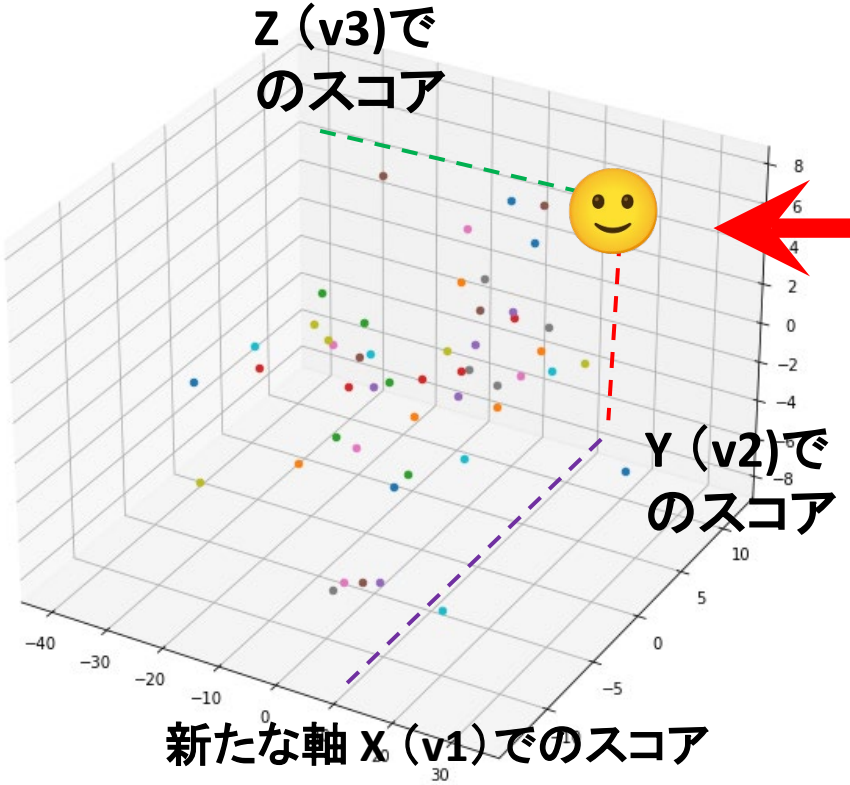
個人の散布図

社会	数学	理科	音楽	美術	体育	技家	英語
43	51	63	60	66	37	44	20
21	49	56	70	72	56	63	16
30	23	57	69	76	33	54	6
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .

8次元の行ベクトル．8個の教科の軸
8次元空間中の新しい軸を
科目を組みあせてつくる

軸に求める基準：
ばらつき最大化

A: 166名 × 8科目
$|Av|^2$ の最大化
v'A'Av を最大化する
8次元ベクトル
（8次元固有ベクトル）



僕、元々8次元．
科目の合成得点
のスコアでスリム
に表現されます．

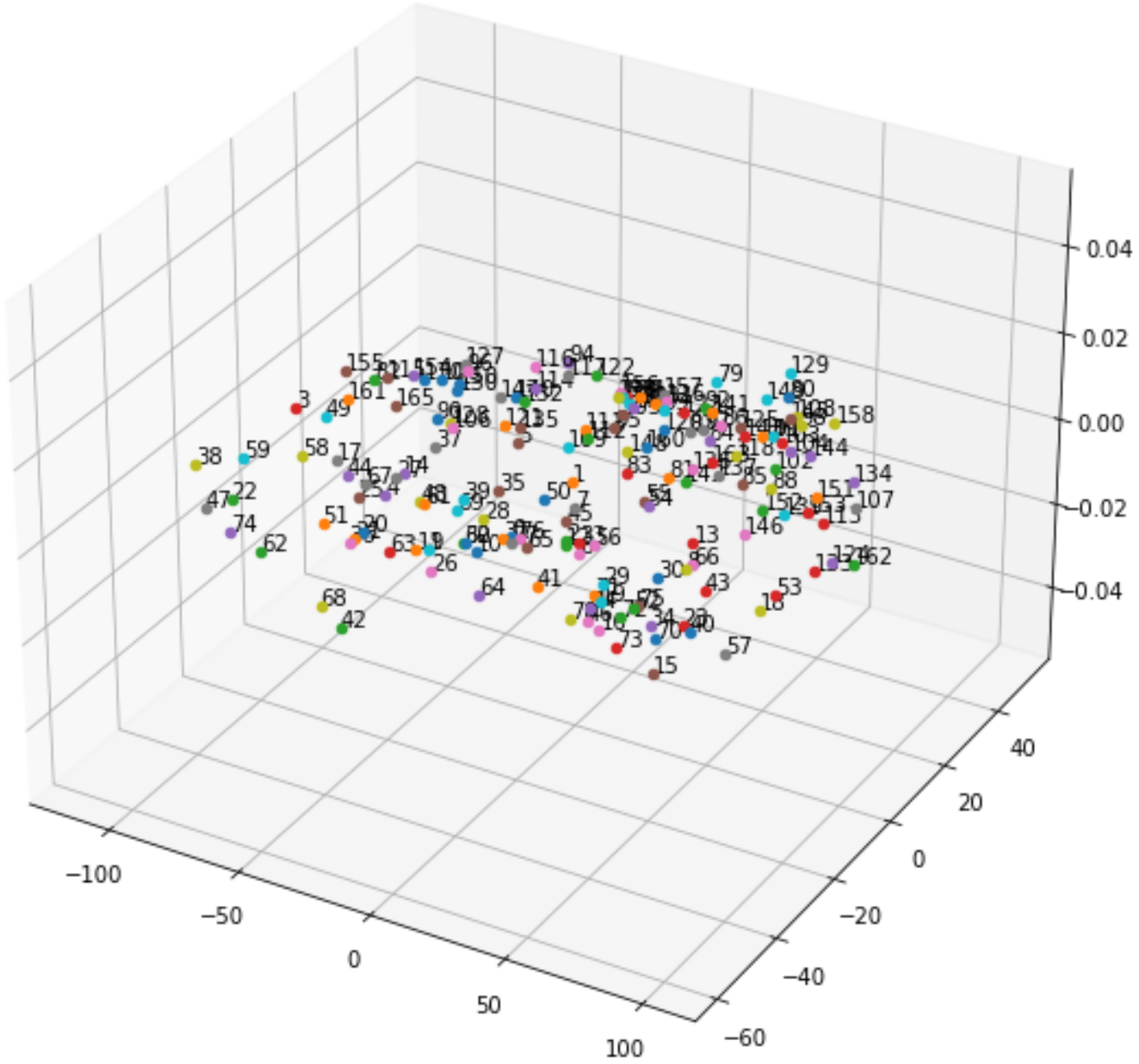
166 x 8 データから作った散布図

あまり良くわからない.

最重要な軸は、ばらつき最大の軸で
　左端: 成績優秀者の塊
　右橋: 下位の生徒の塊
2科目のときと同様.

2番目にバラつきの大きな軸は、
　理系か文系かを示すことがあると、
　書いてあるテキストもある.

この場合は、
　単に 成績にむらがある生徒が
　目についた
　（こうした生徒は主軸では
　　表現できず、第2番目以降の
　スコアーを使わないと
　うまく近似できない）



[73 94 96 77 69 39 74 99]
[60 85 66 35 59 7 52 41]

などは、第2軸（第2番目の固有ベクトル方向）は確かにむらがある.

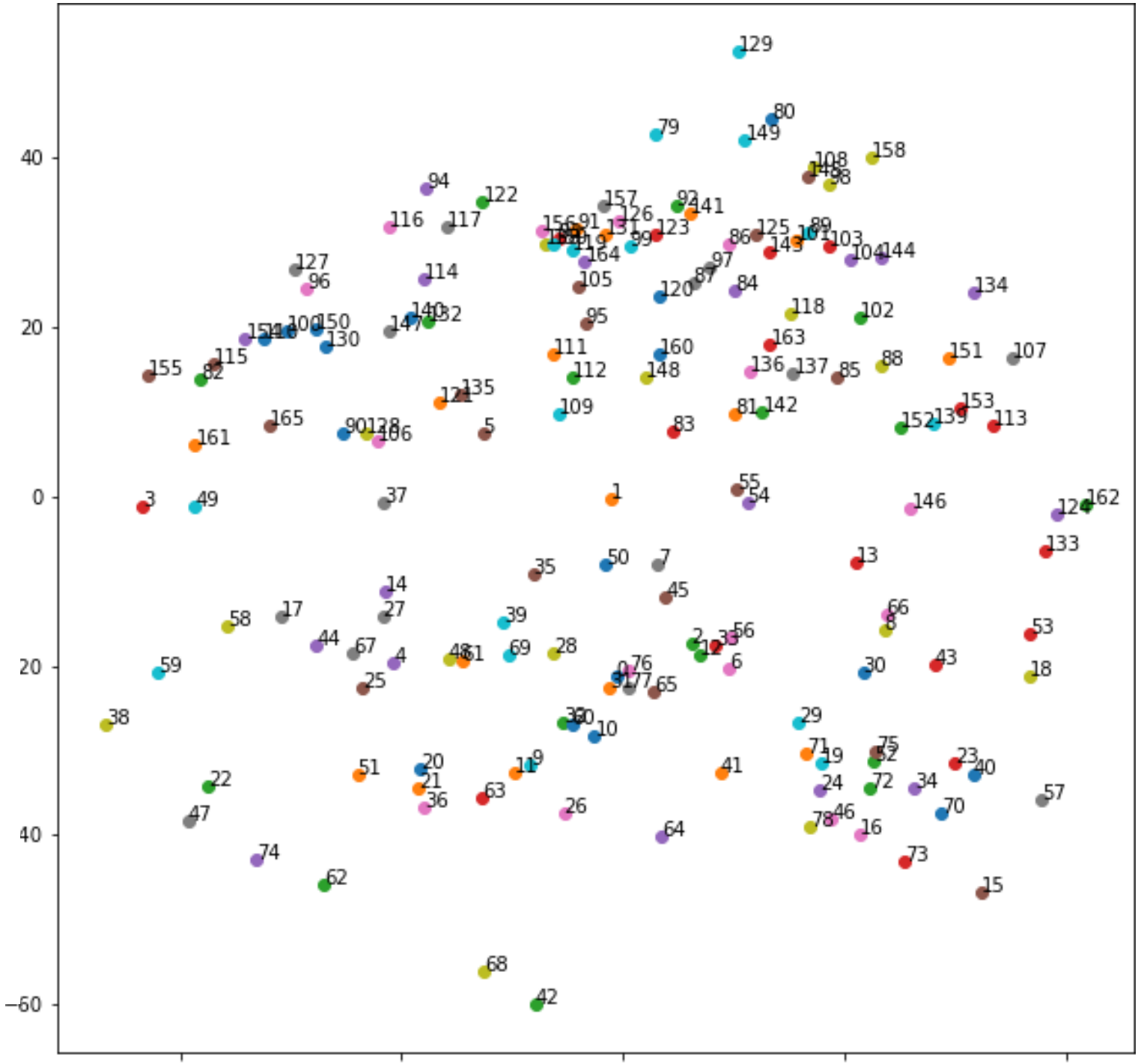
理系・文系 を第2軸が表しているかどうかは？

主軸と第3軸で分析すると、
さらに詳しいことがわかる可能性はある.

一般に、主成分分析は中程度の大きさの固有値は考えない

より細かく分析する場合は、そした固有値の固有ベクトルも観察すべきであろう.

そのスコア—は内積をとるだけなので、（分析は大変だが）計算は瞬時.



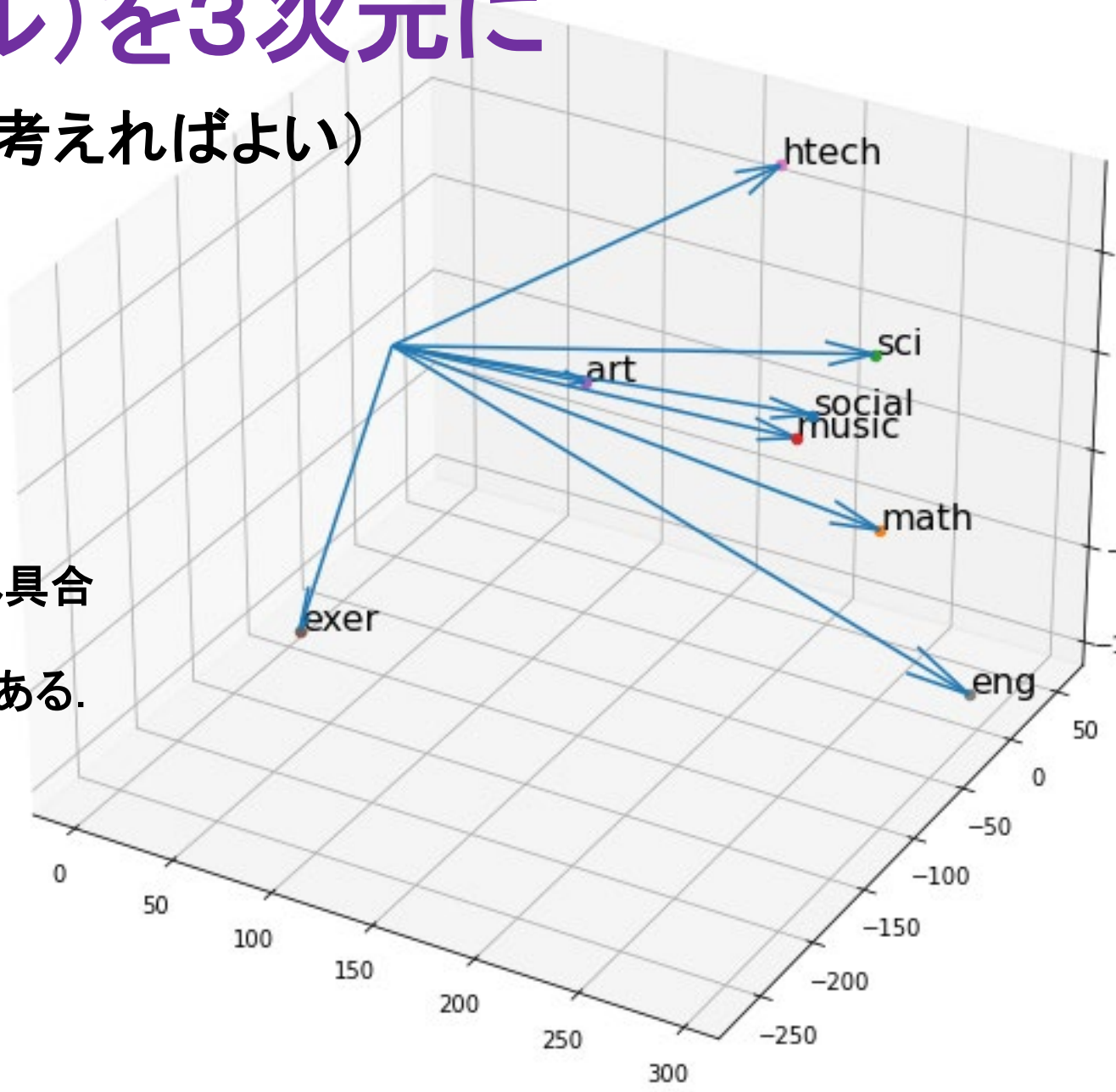
科目（166次元ベクトル）を3次元に

（技術的には単に、転置行列を考えればよい）

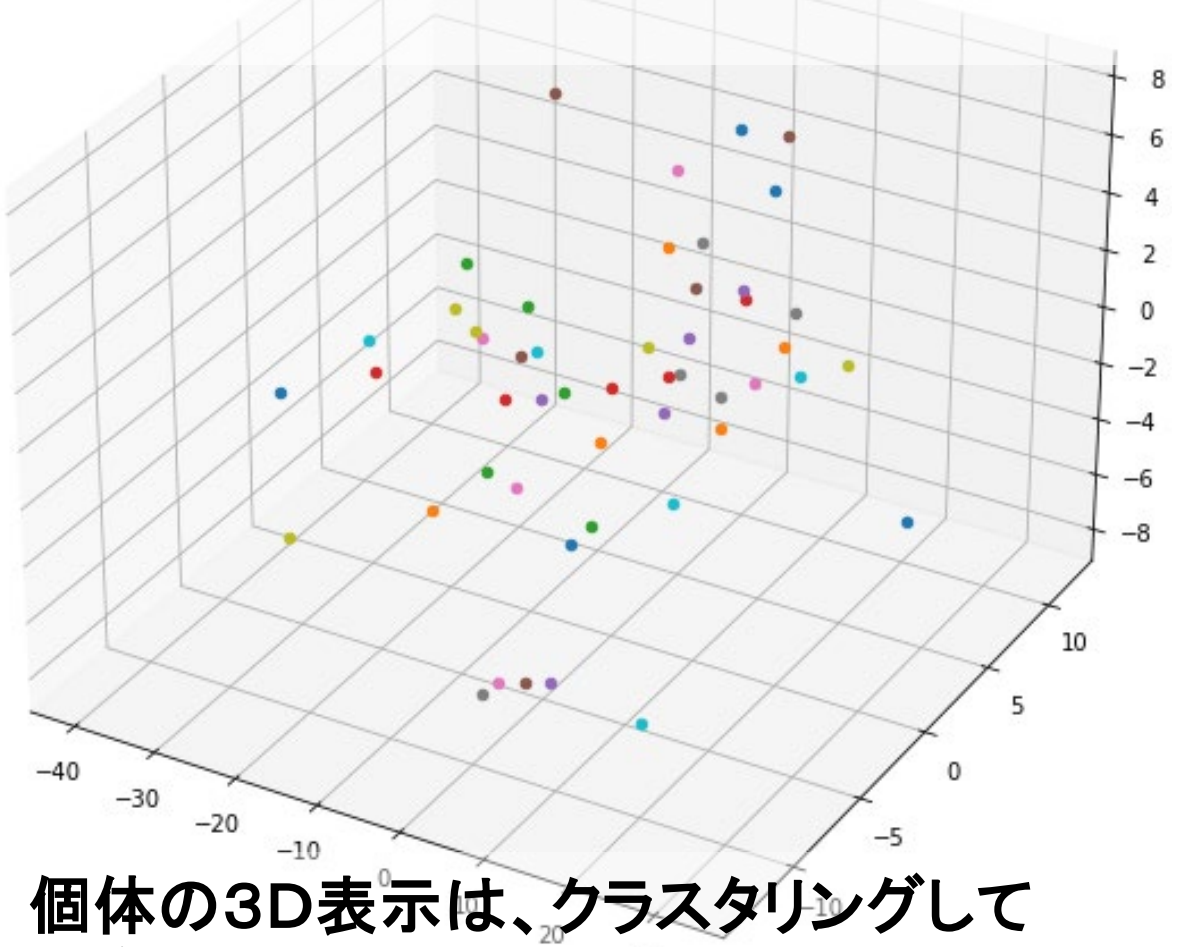
どうだろうか．．．

体育(exer) が他との相関が低いことはわかる

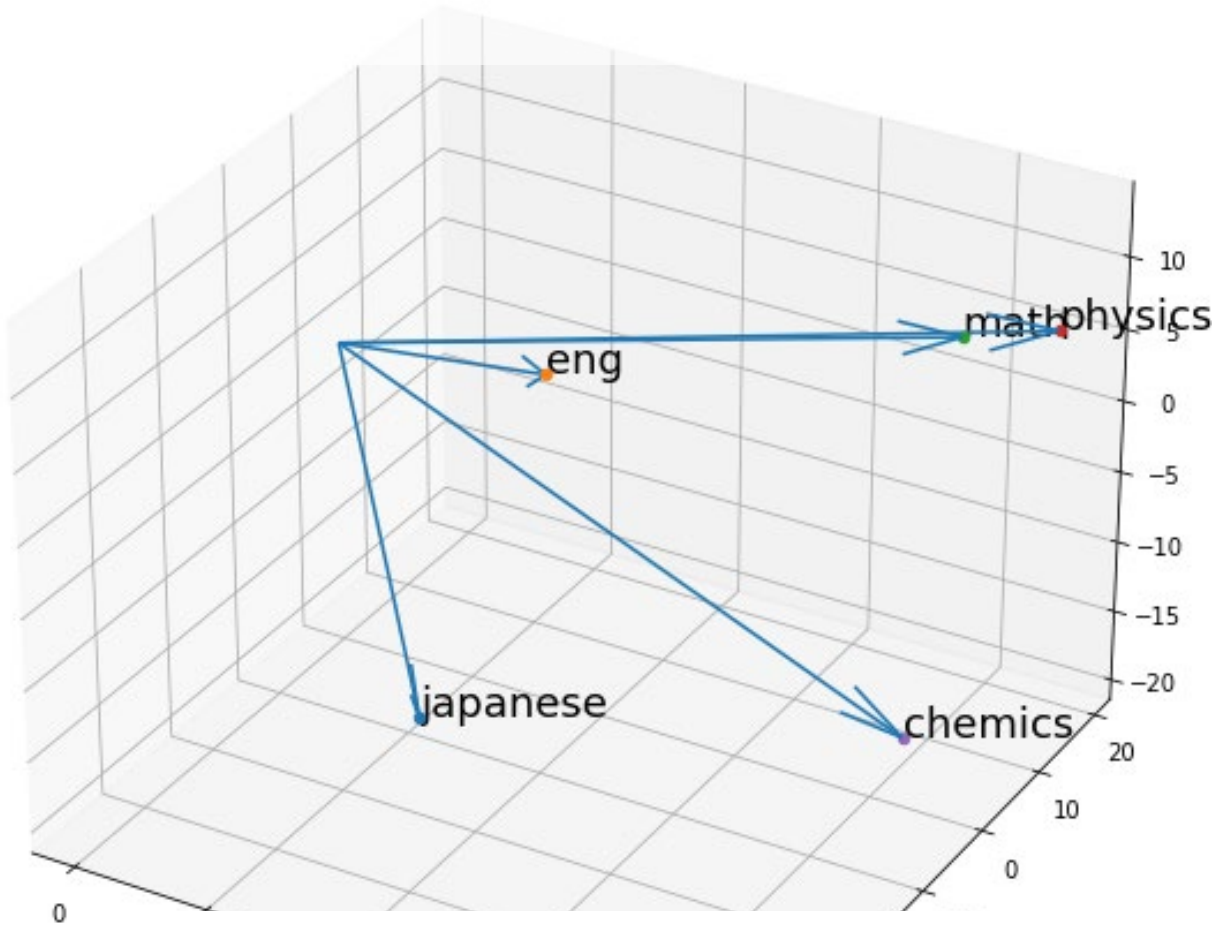
一般に、相関係数をみれば、ペアでの近さ・離れ具合を計量することはできるが、
次元縮約は、近いもの同士を視認できる利点はある．
⇒ クラスタリングの前処理としての次元削減



別の比較的分かりやすいデータ（50x5）



個体の3D表示は、クラスタリングして中身を精査しないとよくわからない．．．（意味不明のクラスターが形成されることも多々ある）



数学と物理はほぼ重なってみえる．
（直感とも合うが．．．）
化学はこうしたものだろうか ？

今回のまとめ

- 次元縮約は他にもいろいろあるが、今回は基礎中の基礎
- 可視化をすることにより、
  直感にあってる、あってない？
  なるほどと思う、あるいは何か変？、
  データがおかしい？
  等々、様々な疑問が生じる.
- **それを契機としてさらなる分析が始まる. . .**