探索的データ解析の現代化

データサイエンスコンソーシアム，慶應義塾大学

柴田　里程

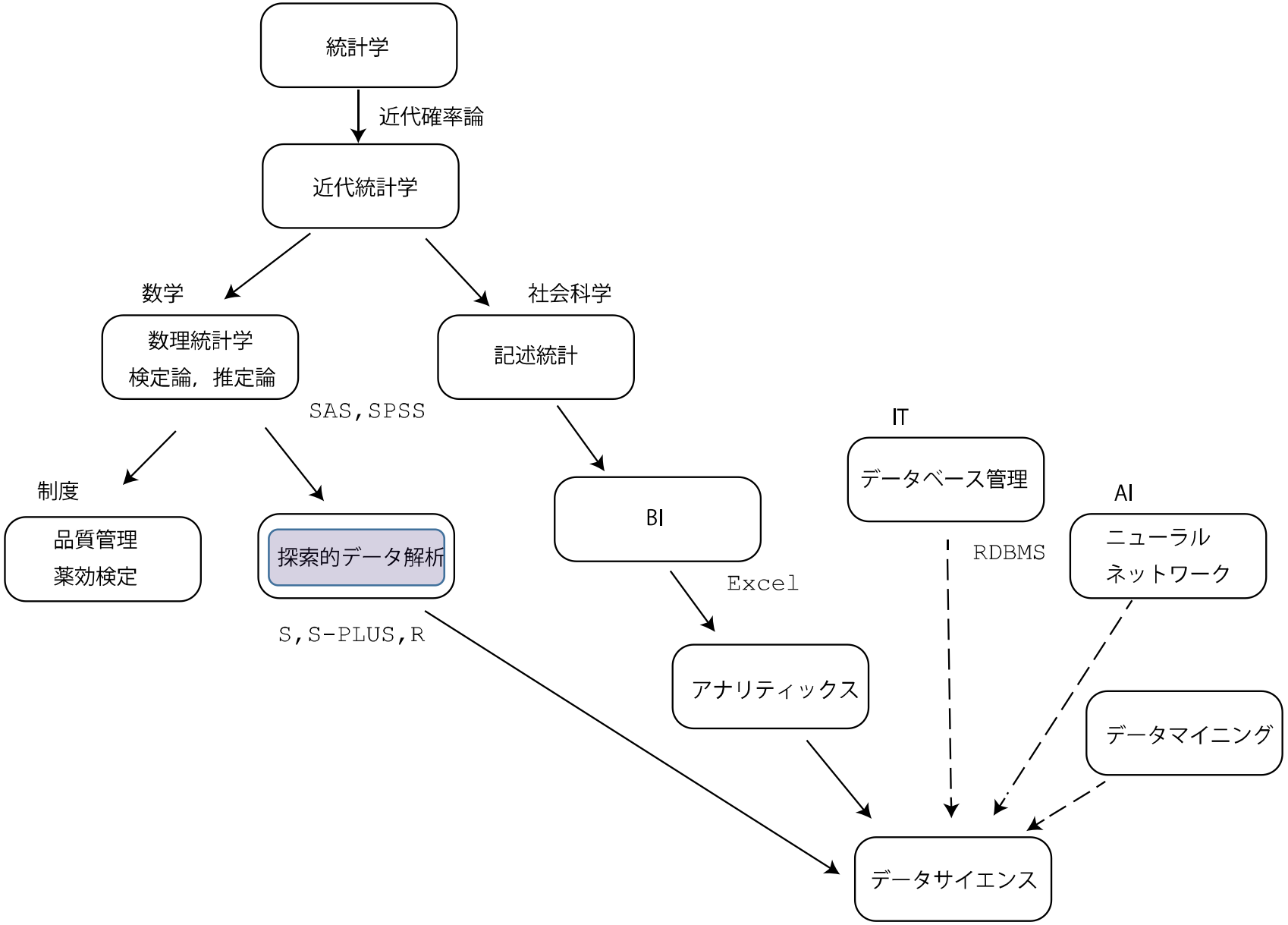
Tukey の時代と現代

EDA

探索的データ解析（EDA）を提唱 (1977)

J. Tukey (1915-2000)
MS: Chemistry from Brown University
DSC: Mathematics from Princeton University
Professor in Princeton University and Research Unit Head in Bell Labs.

Better to have an approximate answer to the right question than a precise answer to the wrong question


統計学
近代確率論
近代統計学
数学
社会科学
数理統計学
検定論，推定論
記述統計
SAS,SPSS
制度
品質管理
薬効検定
探索的データ解析
BI
IT
データベース管理
AI
ニューラル
ネットワーク
RDBMS
Excel
S,S-PLUS,R
アナリティックス
データマイニング
データサイエンス

統計学からデータサイエンスへ

- 方法の科学
  - 制度の補完
  - 方法の蓄積
  - 一定の枠組みでの良し悪しの議論, 自己完結
  - 従属したパラダイム
  - どの方法をつかえばよいの？
- データの科学
  - 発見
  - データの価値評価
  - 実践
  - オープンエンド
  - 独立したパラダイム

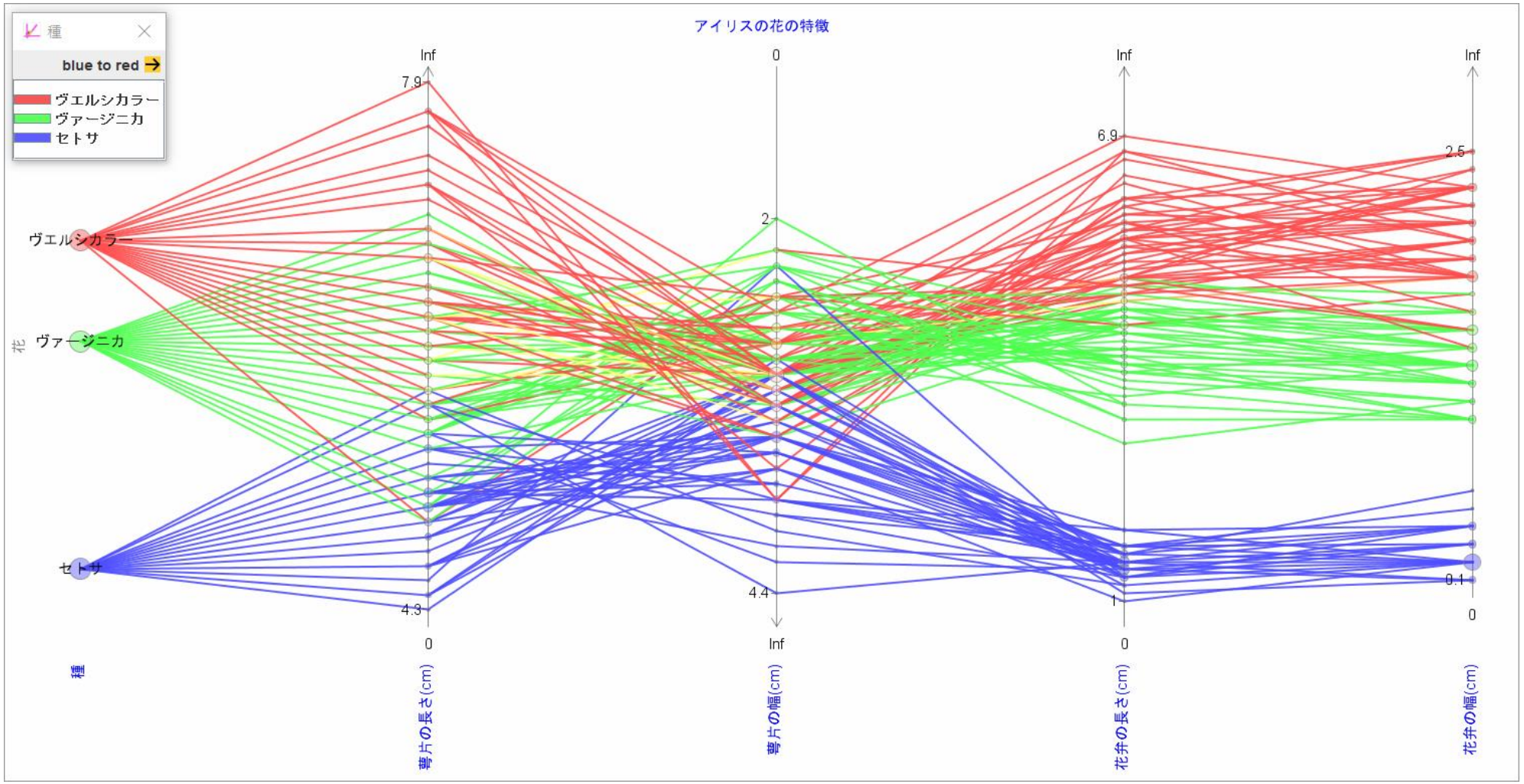
現代化

- 統計学 (EDA も含めて)
  - 幾つかの数値(スカラー)に縮約する(統計量)
  - 幾つかのグラフに縮約する(記述統計)
  - データそのものを理解することは避ける(遠眼鏡)
  - それで？
- 計算能力の制約からの解放
  - 必ずしもいくつかの数値に縮約することにこだわらない
  - データの総体的な理解が可能
    - 本当のデータの価値の発見，評価
    - データの民主化(専門家からの解放)
  - データをもっと身近で，楽しめる存在に
    - Click and Go，GUI

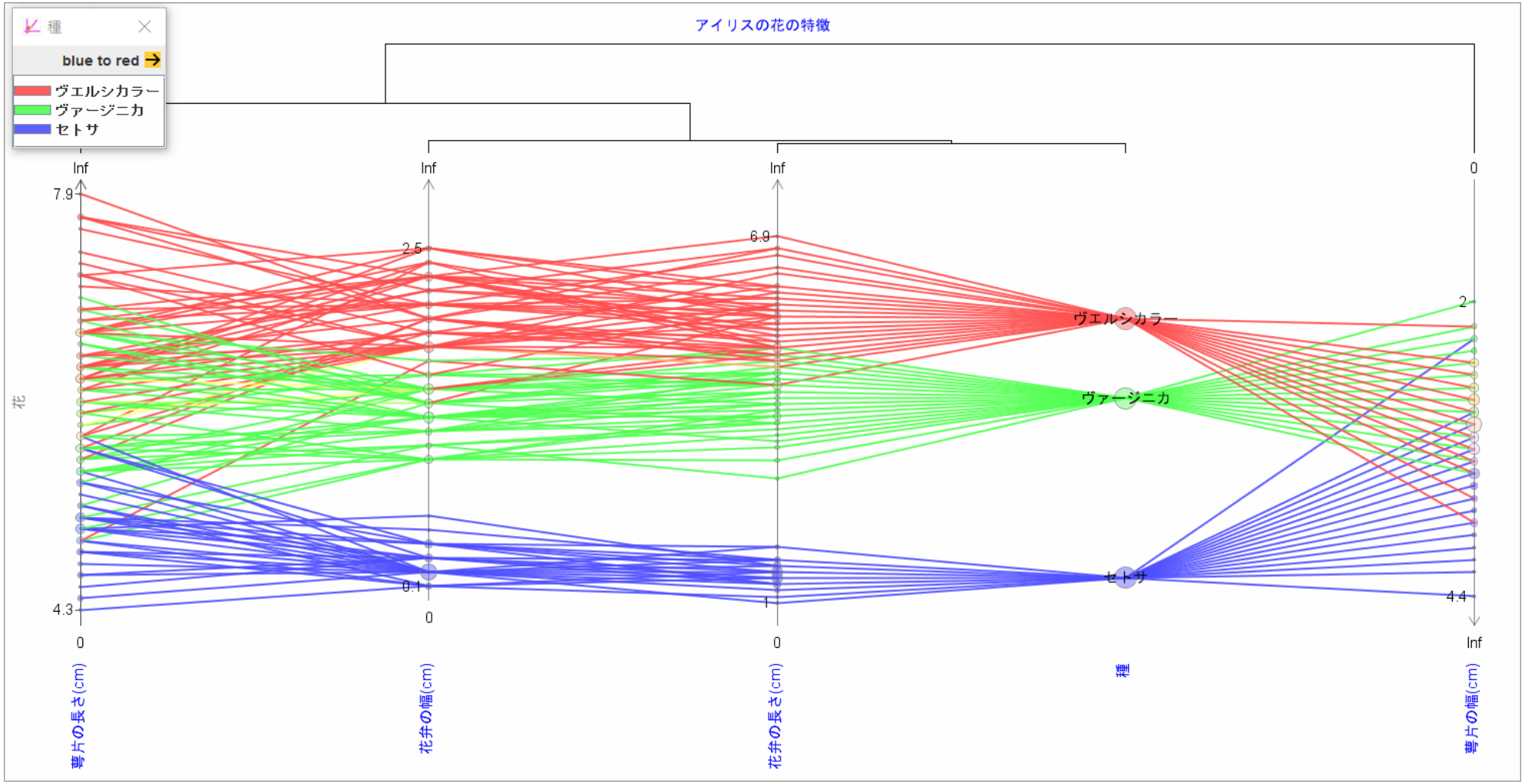
データの価値評価

	A	B	C	D	E	F	G
1		Species	Sepal Length	Sepal Width	Petal Length	Petal Width	
2	1	setosa	5.1	3.5	1.4	0.2	
3	2	setosa	4.9	3	1.4	0.2	
4	3	setosa	4.7	3.2	1.3	0.2	
5	4	setosa	4.6	3.1	1.5	0.2	
6	5	setosa	5	3.6	1.4	0.2	
7	6	setosa	5.4	3.9	1.7	0.4	
8	7	setosa	4.6	3.4	1.4	0.3	
9	8	setosa	5	3.4	1.5	0.2	
10	9	setosa	4.4	2.9	1.4	0.2	
11	10	setosa	4.9	3.1	1.5	0.1	
12	11	setosa	5.4	3.7	1.5	0.2	
13	12	setosa	4.8	3.4	1.6	0.2	
14	13	setosa	4.8	3	1.4	0.1	
15	14	setosa	4.3	3	1.1	0.1	
16	15	setosa	5.8	4	1.2	0.2	
17	16	setosa	5.7	4.4	1.5	0.4	
18	17	setosa	5.4	3.9	1.3	0.4	
19	18	setosa	5.1	3.5	1.4	0.3	
20	19	setosa	5.7	3.8	1.7	0.3	
21	20	setosa	5.1	3.8	1.5	0.3	
22	21	setosa	5.4	3.4	1.7	0.2	
23	22	setosa	5.1	3.7	1.5	0.4	
24	23	setosa	4.6	3.6	1	0.2	
25	24	setosa	5.1	3.3	1.7	0.5	
26	25	setosa	4.8	3.4	1.9	0.2	
27	26	setosa	5	3	1.6	0.2	
28	27	setosa	5	3.4	1.6	0.4	
29	28	setosa	5.2	3.5	1.5	0.2	
30	29	setosa	5.2	3.4	1.4	0.2	

TextilePlot による視覚表現



軸の並べ替え


種
blue to red
ヴエルシカラー
ヴァージニカ
セトサ
アイリスの花の特徴
Inf
7.9
4.3
0
花
Inf
2.5
0.1
0
花弁の幅(cm)
Inf
6.9
1
0
花弁の長さ(cm)
ヴエルシカラー
ヴァージニカ
セトサ
種
0
2
4.4
Inf
萼片の幅(cm)
萼片の長さ(cm)

データ総体の視覚表現

- 直感を呼びさます（入口）
- 特定の見方からの解放
- すばやいデータの価値評価（ゴミかどうかの判断）
- 記録と変量の両面からクロスチェック
- 必要に応じた細部のチェック
- 他人への説明 (出口）

データに〇〇分析を適用したら結果はこうなりました．そうですか・・・

2 階から目薬

○○分析よさようなら

線形モデル（多変量解析）を例に

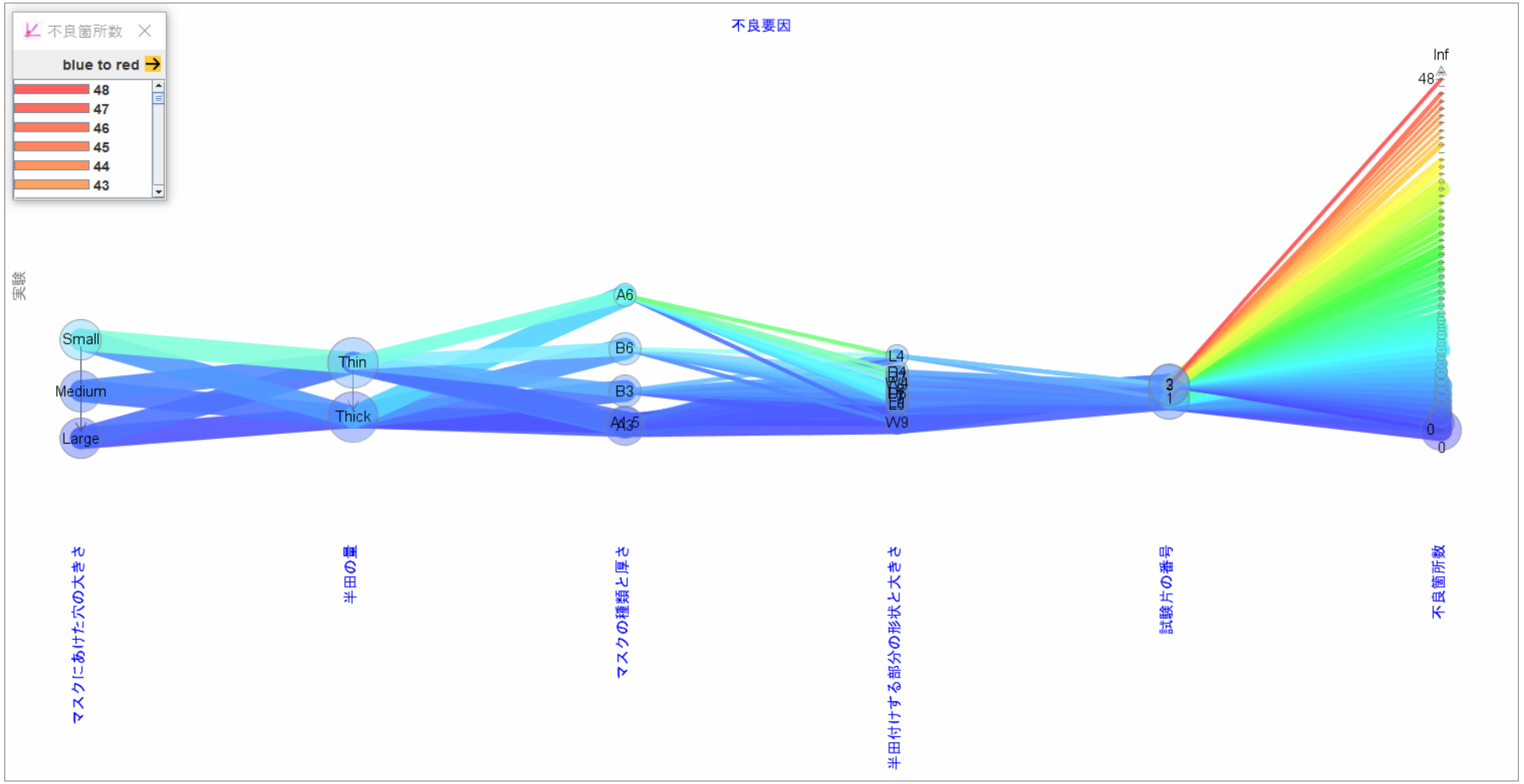
- 回帰分析
- 分散分析
- ロジット分析
- コレスポンデンス分析
- 正準相関分析

- 主成分分析

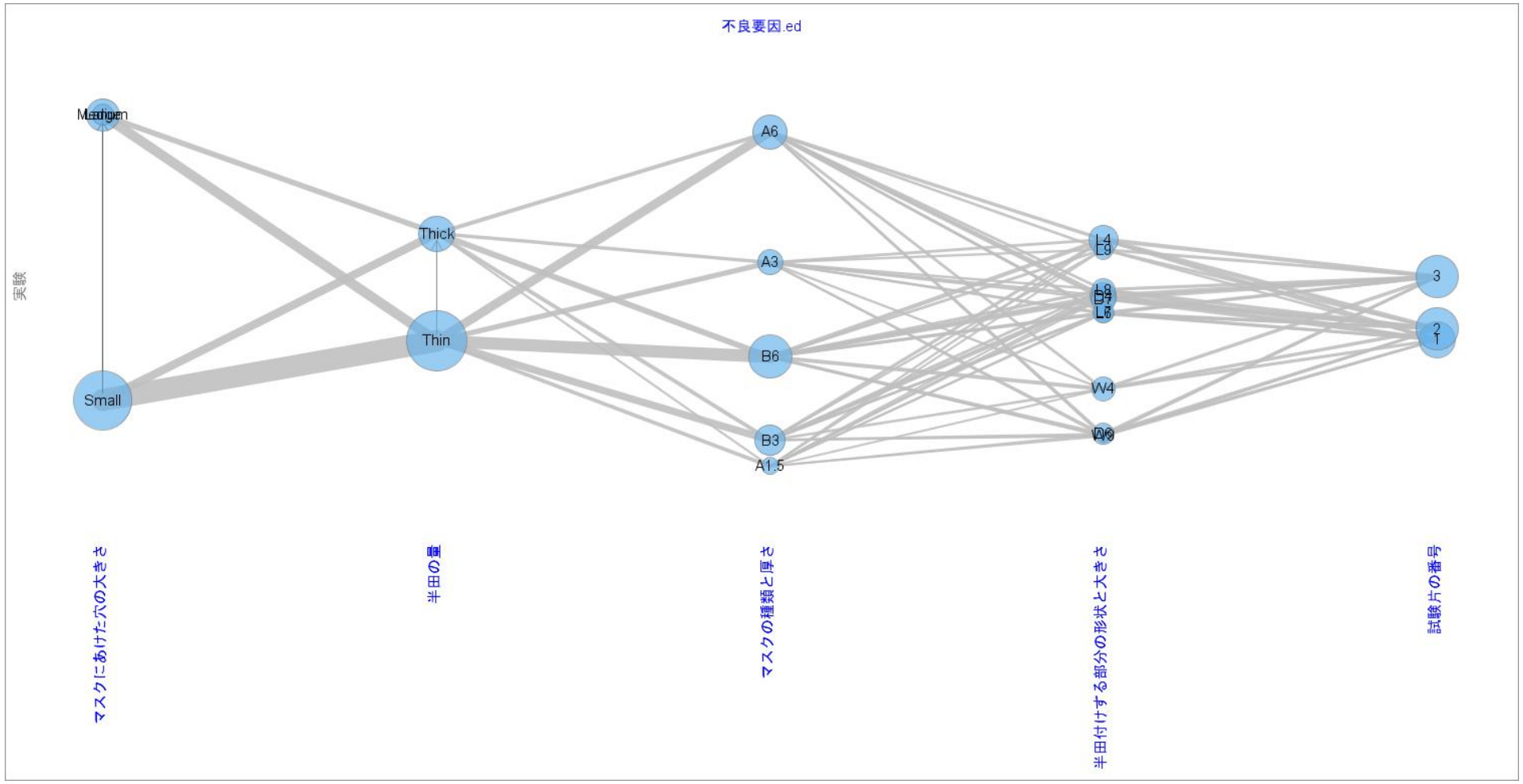
○○分析とさよならするためには

- カテゴリカルデータを別扱いしない
  - 対比による数値化で同等に扱える
  - 被説明変量でさえも
- ベクトルの基本概念を踏まえる
  - ノルム
  - 直交
  - 線形空間
  - 射影

例：プリント基板のはんだ付け実験データ

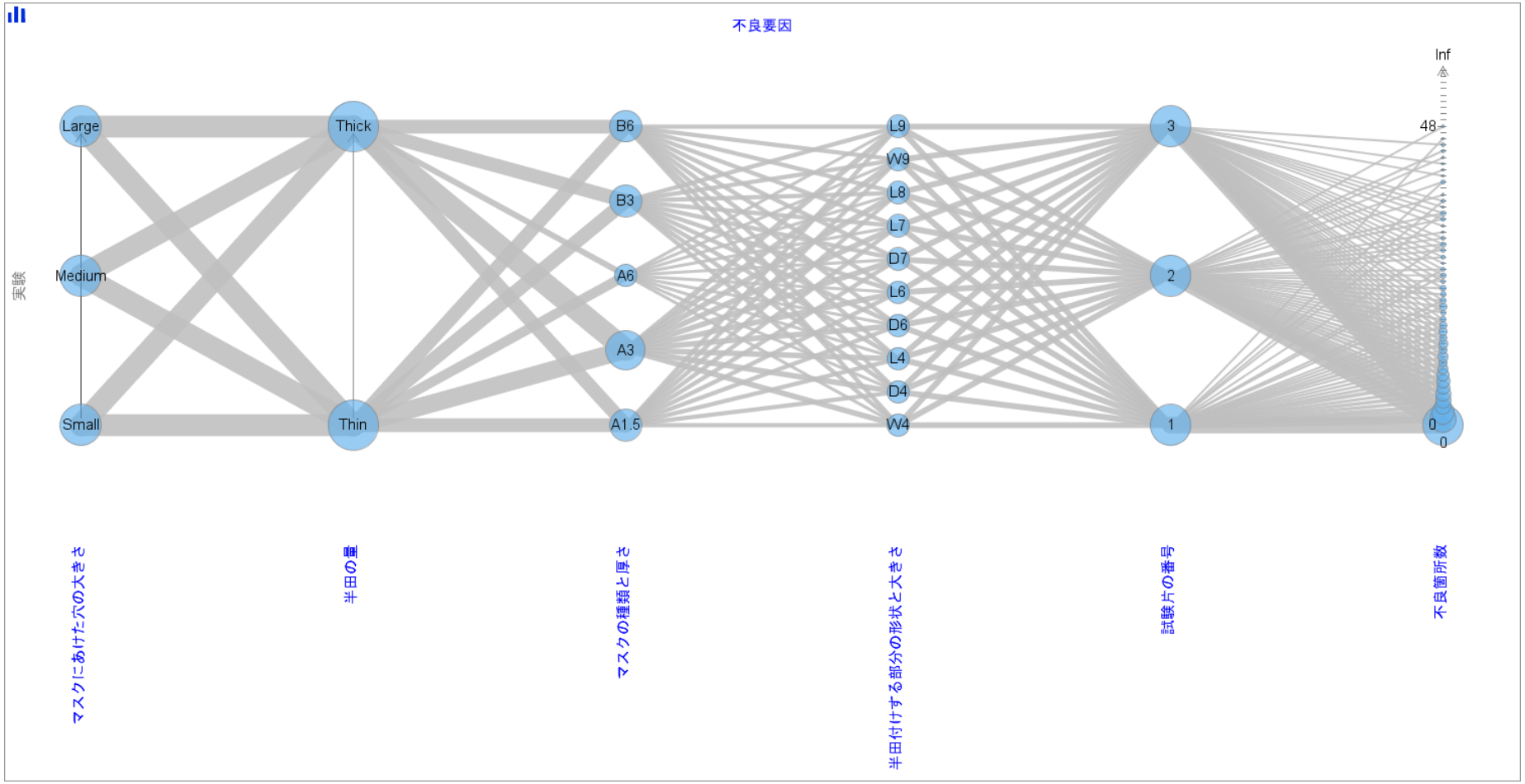


不良箇所数を Weft の太さに反映(高次元分割表)

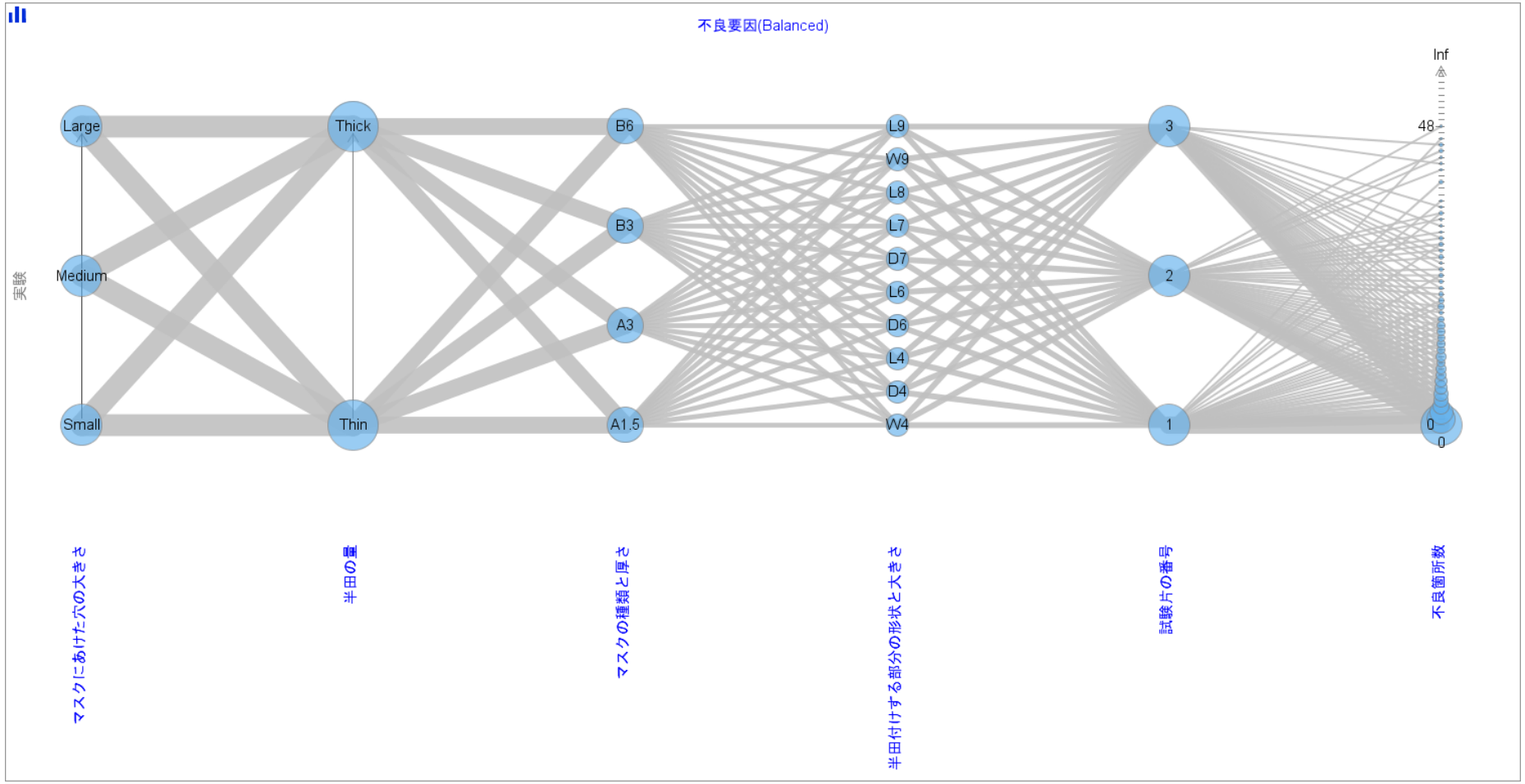


バランスした実験かどうか

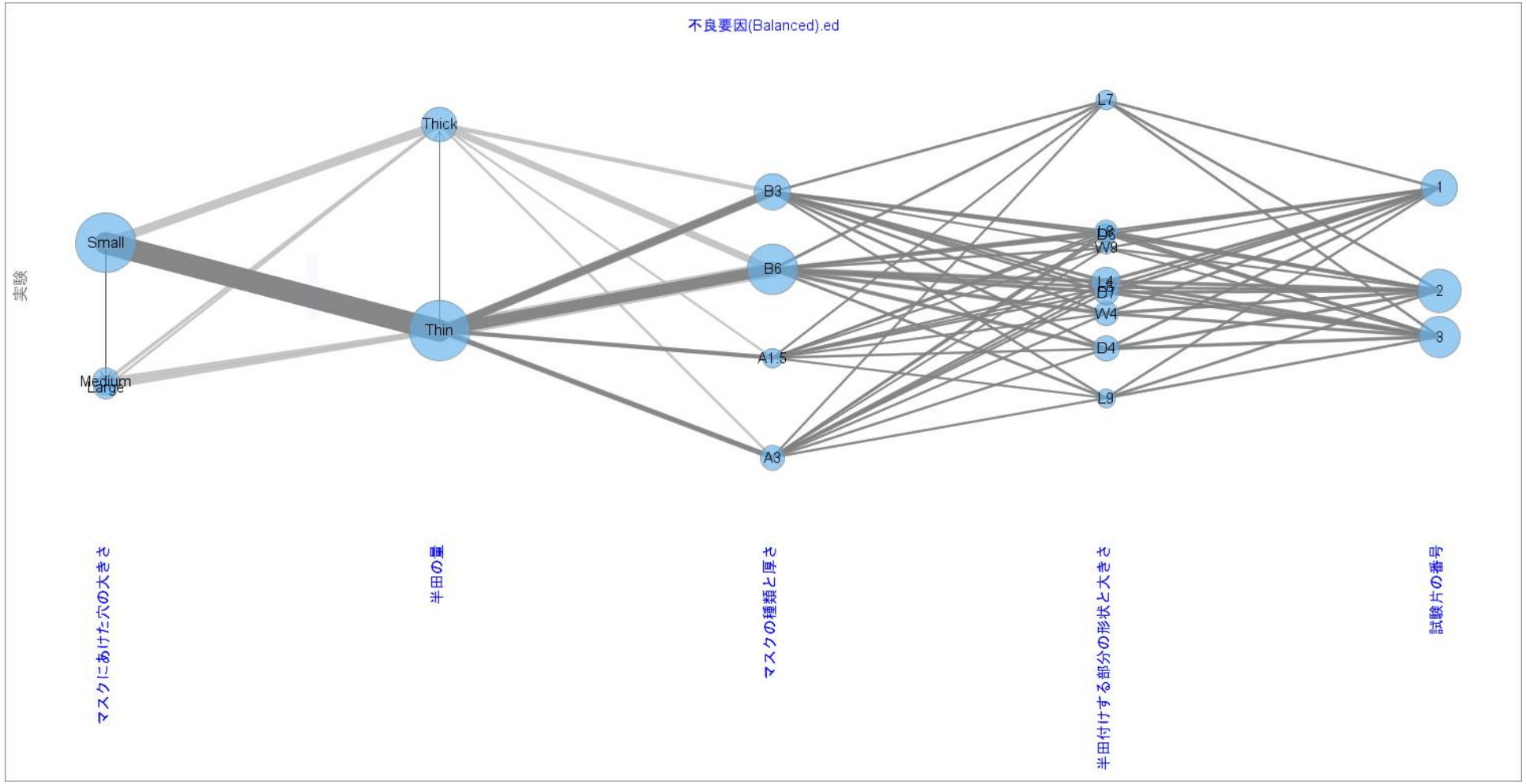
(Weft の太さに実験回数を反映した pcp)



間違えた実験を削除してバランスするように



再び不良箇所数を Weft の太さに反映



すでに大まかな結論は得られた

- 不良が多発する要因の組み合わせ
  - 穴が小さい
  - 半田が薄い
  - マスクのタイプが B3 または B6
- チェンバース・ヘイスティ―『Sと統計モデル』ではここに至るまでに 1 章を費やしている
- ここまででも，不良を少なくする手がかりは得られたので，十分なことも多い
- 更に深堀りするとしたら，回帰分析，分散分析など

R での回帰分析

```
> lm.result=lm(skips~., data=balanced)

> lm.result

Call:
lm(formula = skips ~ ., data = balanced)

Coefficients:
(Intercept)   sizeMedium    sizeLarge  amountThick        maskA3        maskB3
maskB6         padD4
   10.28472      -8.91250      -9.40417      -4.96944       0.86111       3.75000
8.80556       0.69444
      padL4         padD6         padL6         padD7         padL7         padL8
padw9         padL9
    2.69444      -1.36111      -2.55556       0.06944      -1.88889      -0.88889      -
4.38889      -2.44444
     panel2        panel3
    1.60000       1.17083
```

係数の値をみて何がわかる？

回帰係数が有意かどうかは重大なこと？

```
> summary(lm.result)

Call:
lm(formula = skips ~ ., data =
balanced)

Residuals:
    Min      1Q  Median      3Q
Max
-9.0417 -3.4267 -0.7472  2.5021
27.3097

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.28472    0.81844  12.566  < 2e-16 ***
sizeMedium  -8.91250    0.47253 -18.861  < 2e-16 ***
sizeLarge   -9.40417    0.47253 -19.902  < 2e-16 ***
amountThick -4.96944    0.38582 -12.880  < 2e-16 ***
maskA3       0.86111    0.54563   1.578 0.114969
maskB3       3.75000    0.54563   6.873 1.39e-11 ***
maskB6       8.80556    0.54563  16.138  < 2e-16 ***
padD4        0.69444    0.86271   0.805 0.421118
padL4        2.69444    0.86271   3.123 0.001862 **
padD6       -1.36111    0.86271  -1.578 0.115083
padL6       -2.55556    0.86271  -2.962 0.003157 **
padD7        0.06944    0.86271   0.080 0.935866
padL7       -1.88889    0.86271  -2.189 0.028891 *
padL8       -0.88889    0.86271  -1.030 0.303204
padW9       -4.38889    0.86271  -5.087 4.67e-07 ***
padL9       -2.44444    0.86271  -2.833 0.004737 **
panel2       1.60000    0.47253   3.386 0.000749 ***
panel3       1.17083    0.47253   2.478 0.013453 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.176 on 702 degrees of freedom
Multiple R-squared:  0.6091,  Adjusted R-squared:  0.5996
F-statistic: 64.34 on 17 and 702 DF,  p-value: < 2.2e-16
```

回帰係数は直交射影の係数に過ぎない

$$\mathbf{y} = \underline{\hat{\beta}_0 \mathbf{1} + \hat{\beta}_1 \mathbf{x}_1 + \cdots + \hat{\beta}_p \mathbf{x}_p} + \hat{\varepsilon}$$

$\mathbf{y}$ の Span($\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p$) への直交射影

R での分散分析

```
> anova(lm.result)
Analysis of Variance Table

Response: skips
           Df  Sum Sq Mean Sq  F value    Pr(>F)
size        2 13449.0  6724.5 250.9723 < 2.2e-16 ***
amount      1  4445.2  4445.2 165.9027 < 2.2e-16 ***
mask        3  8521.2  2840.4 106.0097 < 2.2e-16 ***
pad         9  2562.3   284.7  10.6254 1.559e-15 ***
panel       2   329.2   164.6   6.1433  0.002265 **
Residuals 702 18809.3    26.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1
```

$$\mathbf{y} = \hat{\beta}_0\mathbf{1} + \hat{\beta}_1\mathbf{x}_1^* + \cdots + \hat{\beta}_p\mathbf{x}_p^* + \hat{\varepsilon}$$

$\{\mathbf{1}, \mathbf{x}_1^*, \ldots, \mathbf{x}_p^*\}$ は $\{\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_p\}$ の直交化

$$\| \mathbf{y} \|^2 = \| \hat{\beta}_0\mathbf{1} \|^2 + \| \hat{\beta}_1\mathbf{x}_1^* \|^2 + \cdots + \| \hat{\beta}_p\mathbf{x}_p^* \|^2 + \| \hat{\varepsilon} \|^2$$

ピタゴラス

回帰分析と分散分析

- 違い
  - 説明変量ベクトルを直交化するかどうか
  - 係数に注目するか，射影ベクトルそのものに注目するか

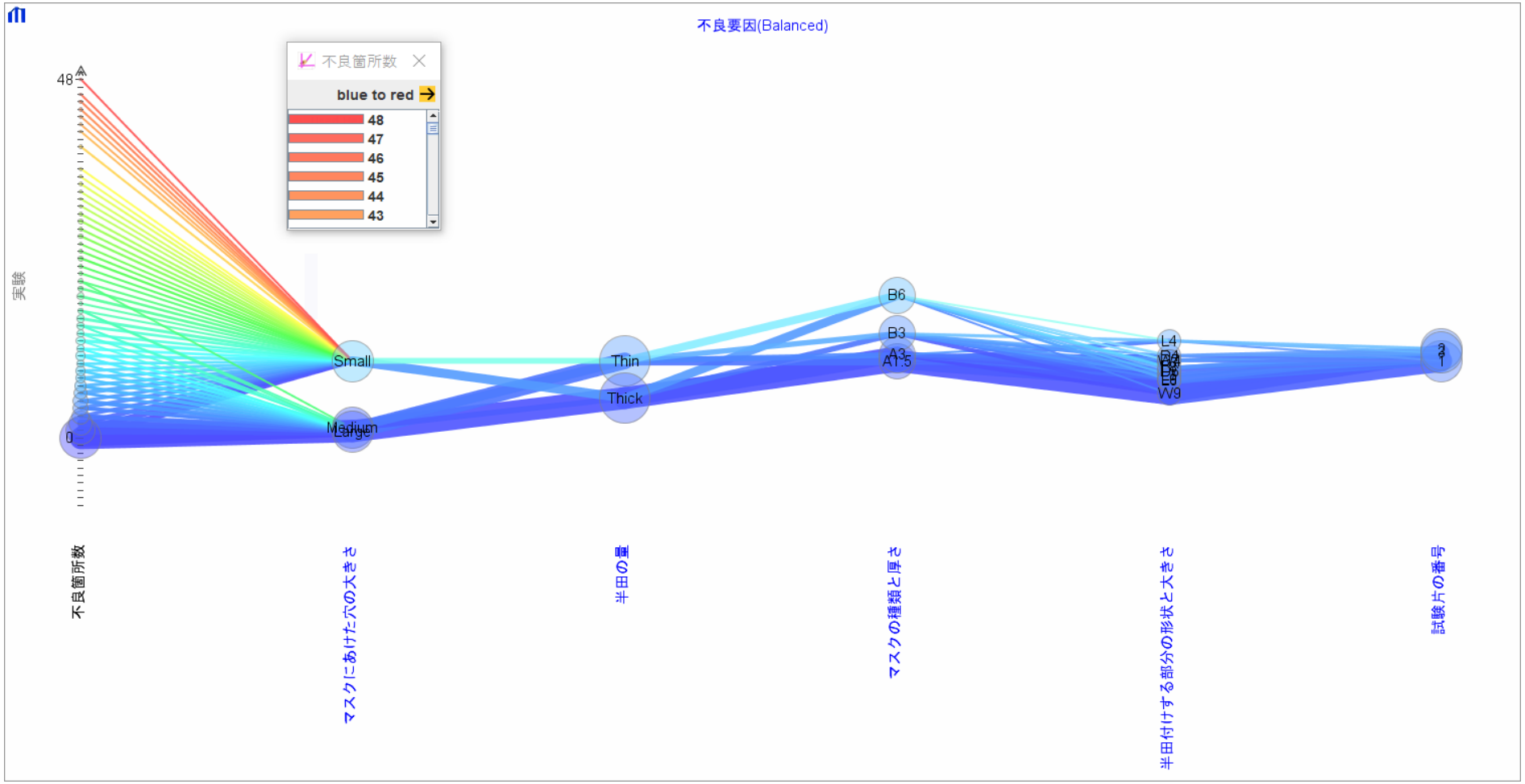
$\mathbf{y}$ と $\hat{\beta}_0\mathbf{1}, \hat{\beta}_1\mathbf{x}_1, \cdots, \hat{\beta}_p\mathbf{x}_p$

あるいは

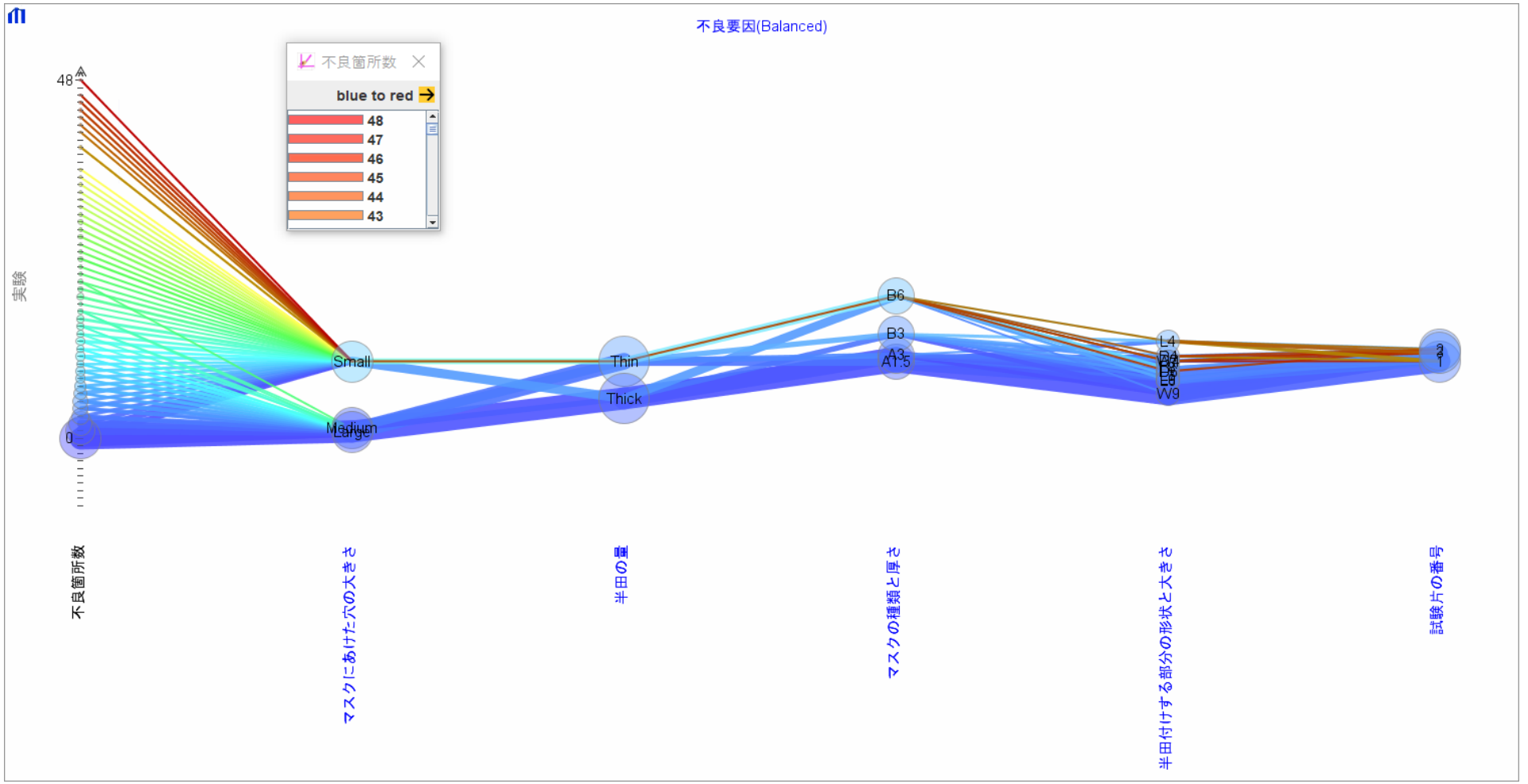
$\mathbf{y}$ と $\hat{\beta}_0\mathbf{1}, \hat{\beta}_1\mathbf{x}_1^*, \cdots, \hat{\beta}_p\mathbf{x}_p^*$

の関係をその大きさとともに調べればよい

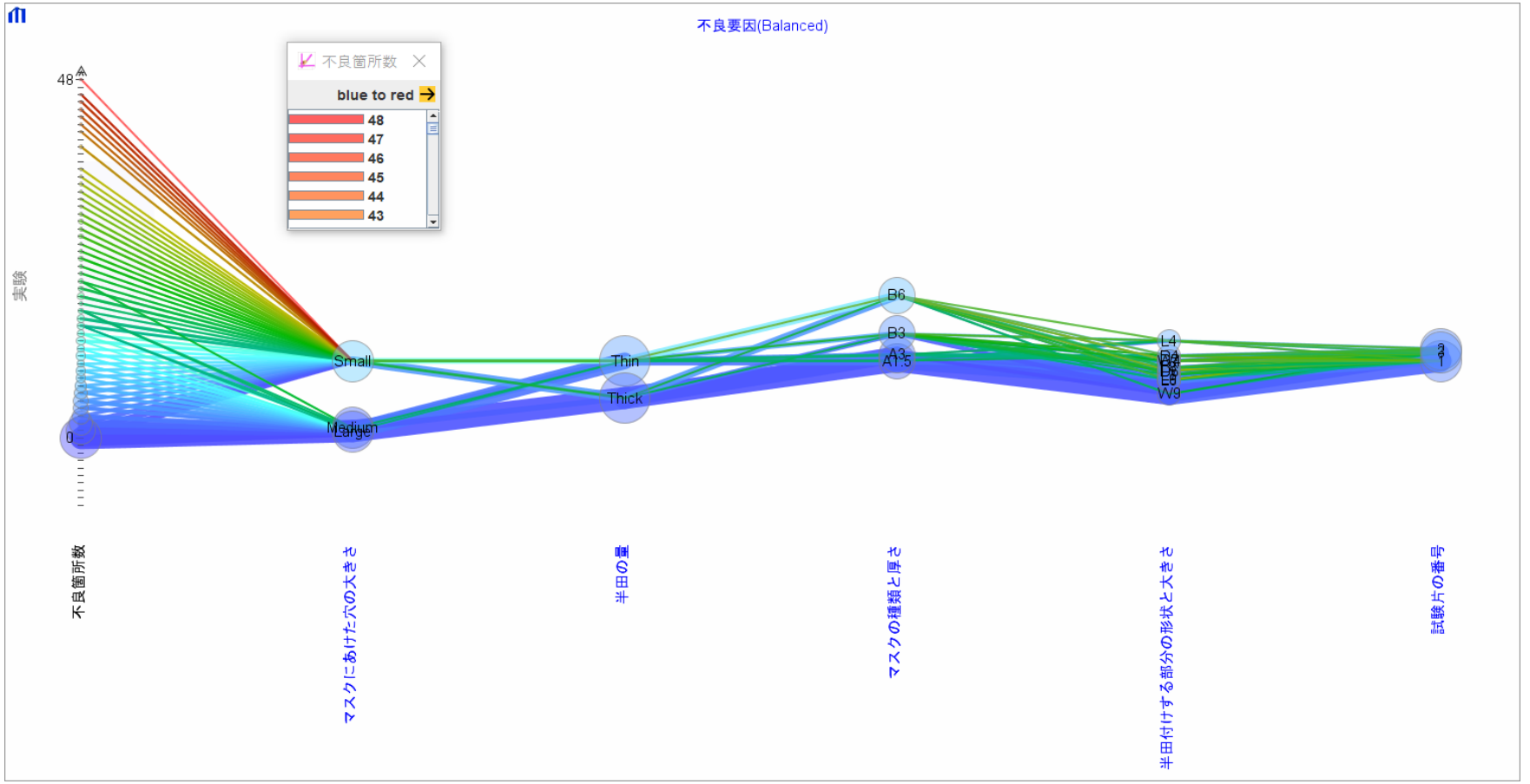
$\mathbf{y}$ と $\hat{\beta}_0\mathbf{1}, \hat{\beta}_1\mathbf{x}_1, \cdots, \hat{\beta}_p\mathbf{x}_p$ の関係



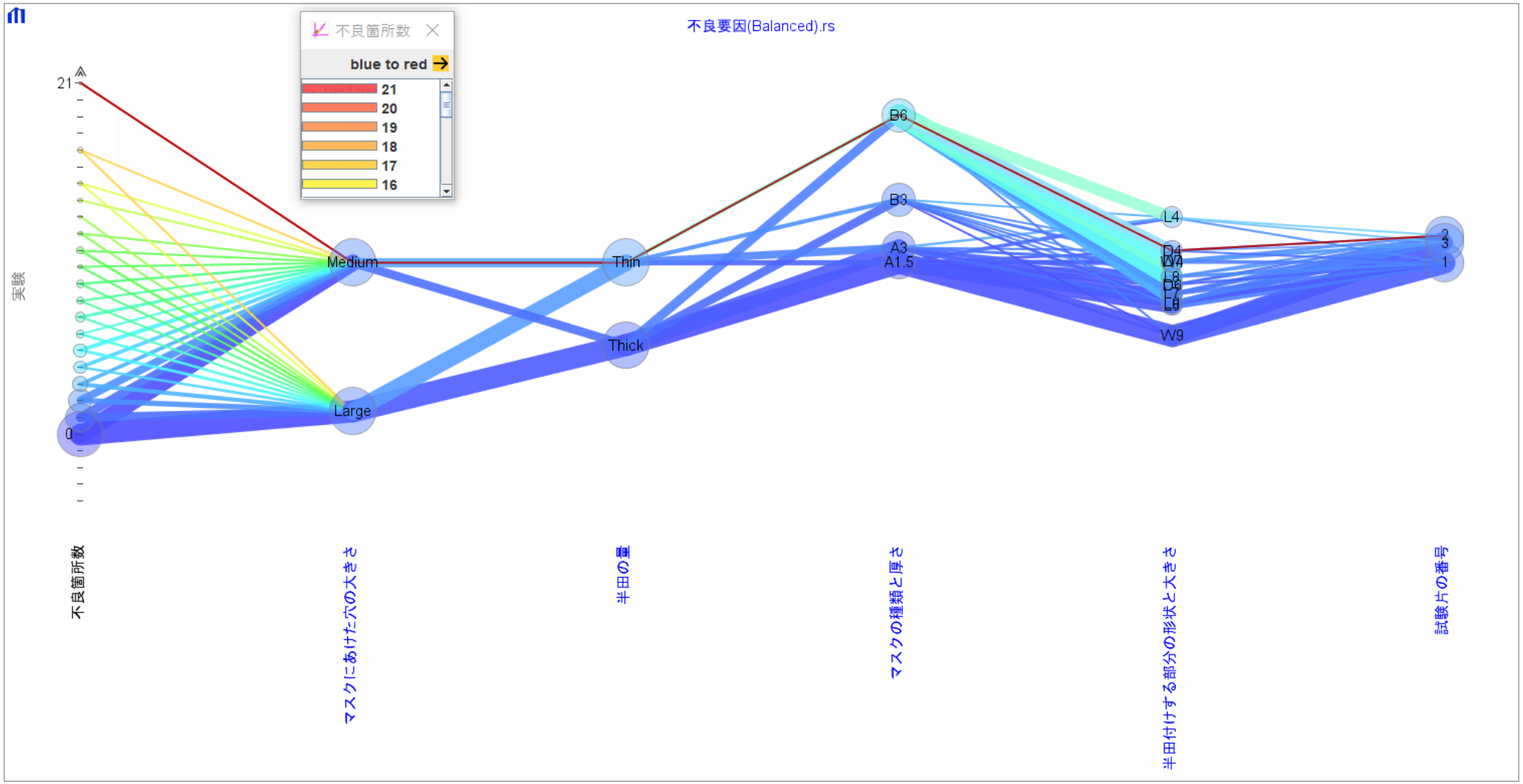
不良箇所数の多い実験


不良要因(Balanced)
不良箇所数
blue to red
48
47
46
45
44
43
実験
48
0
Small
Medium
Large
Thin
Thick
B6
B3
A3
A1
L4
W9
不良箇所数
マスクにあけた穴の大きさ
半田の量
マスクの種類と厚さ
半田付けする部分の形状と大きさ
試験片の番号

不良箇所数の比較的多い実験



穴の大きさが small の実験を除いたら



実装にあたって検討すべきこと

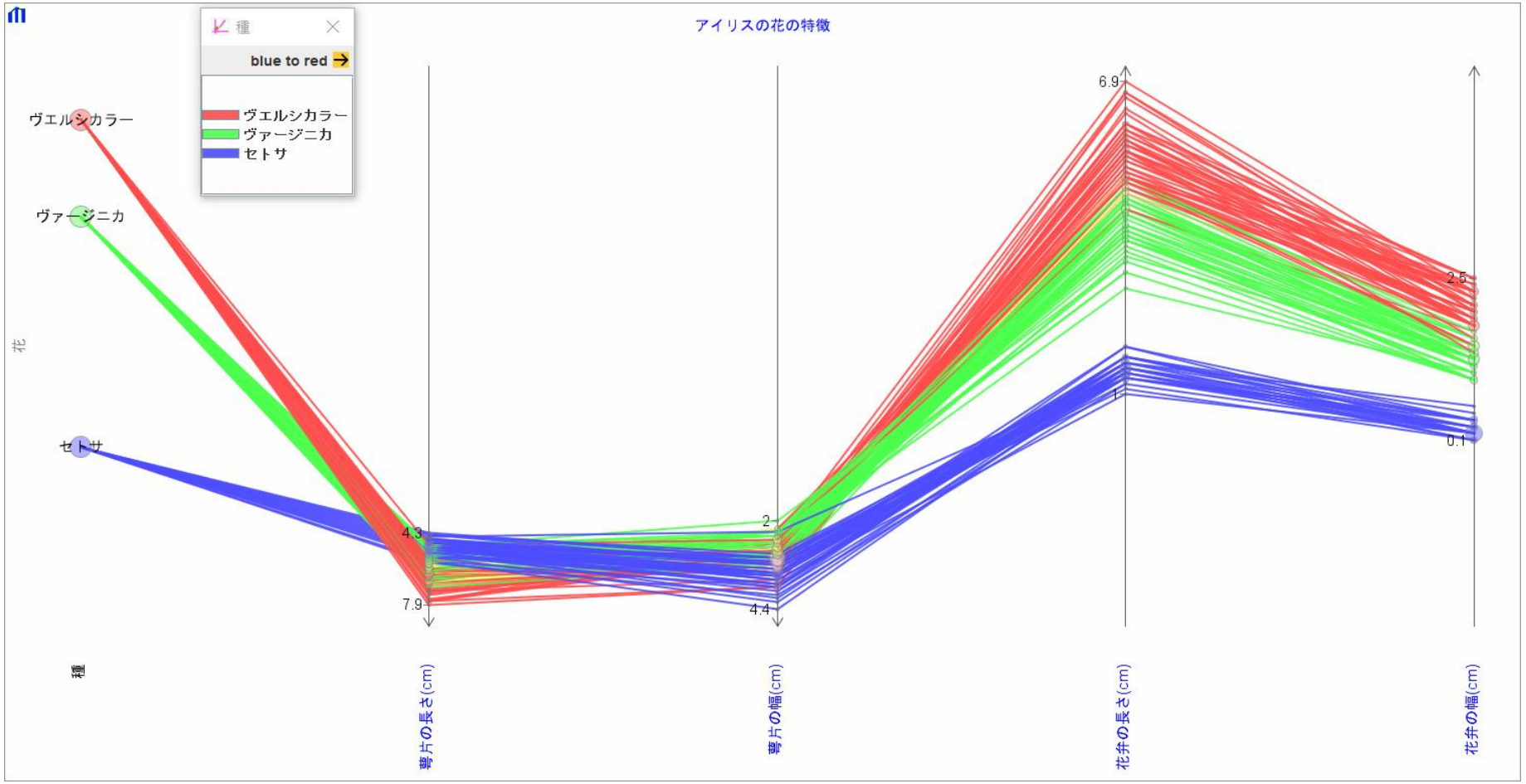
- あてはめを R から行うか TAD から行うか(ユーザインタフェース)
  - データマニュピレーションによる再あてはめ
- $\mathbf{y}$ と $\hat{\beta}_0\mathbf{1}, \hat{\beta}_1\mathbf{x}_1^*, \cdots, \hat{\beta}_p\mathbf{x}_p^*$ の表示
  - 直交化の仕方と解釈
- 説明変量の部分和
- 交互作用
  - 特にカテゴリカルとニューメリカルの交互作用
    - 型が混合したデータベクトルの表示

被説明変量がカテゴリカル

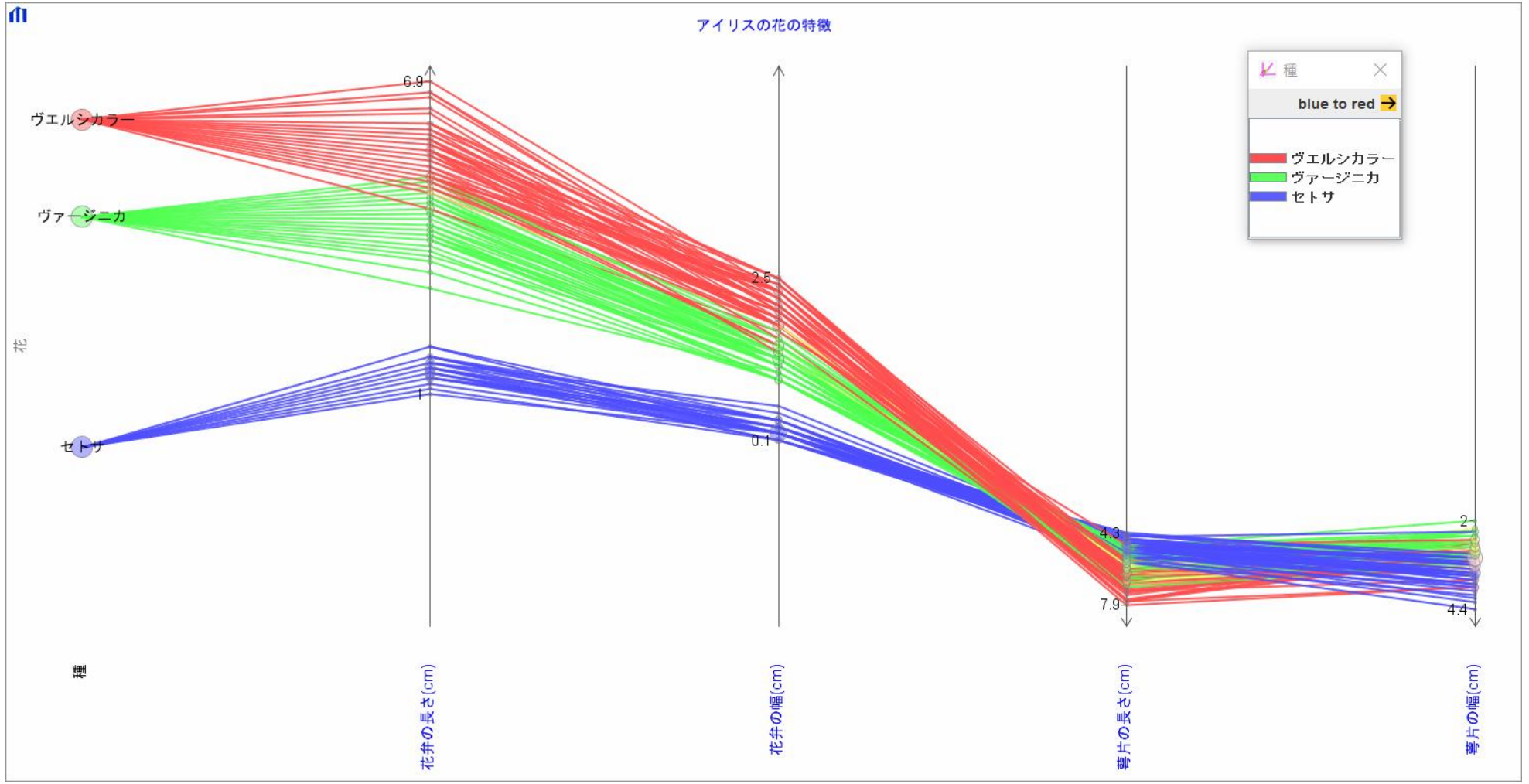
- ロジットモデルあるいは多項ロジット（ロジスティック）モデル
  - 水準の確率のモデル化であって，水準そのもののモデルではない
  - 結果はわかったようでわからない．直感的な理解は困難
  - 多項ロジットになるとパラメータが各水準に対して(説明変量数+1)個導入され，パラメータを介しての理解は困難
- 被説明変量も対比でコーディングした上での正準相関による線形モデルの当てはめ

$$\| Y\mathbf{a} - X\mathbf{b} \|^2 \to \min \ \ (\ \| Y\mathbf{a} \| = \| \ X\mathbf{b} \| = 1 \ )$$

正準相関による当てはめ結果



軸の順序の入れ替え


アイリスの花の特徴
種
blue to red
ヴエルシカラー
ヴァージニカ
セトサ
ヴエルシカラー
ヴァージニカ
セトサ
花
6.9
2.5
1
0.1
4.3
7.9
2
4.4
種
花弁の長さ(cm)
花弁の幅(cm)
萼片の長さ(cm)
萼片の幅(cm)

まとめ

- 方法の科学からの脱却
  - 無責任
  - 消極的
- 少数の縮約値に落とすパラダイムとの決別
  - 時代の反映
  - データの大量複雑化
  - 独立したパラダイム
- データを総体として眺めれば，ある程度の結論は得られてしまう
  - データを眺めるために必要となる理論研究とソフトウエア開発
- その上で，必要ならモデルの当てはめによって深堀りする