TRADのいま

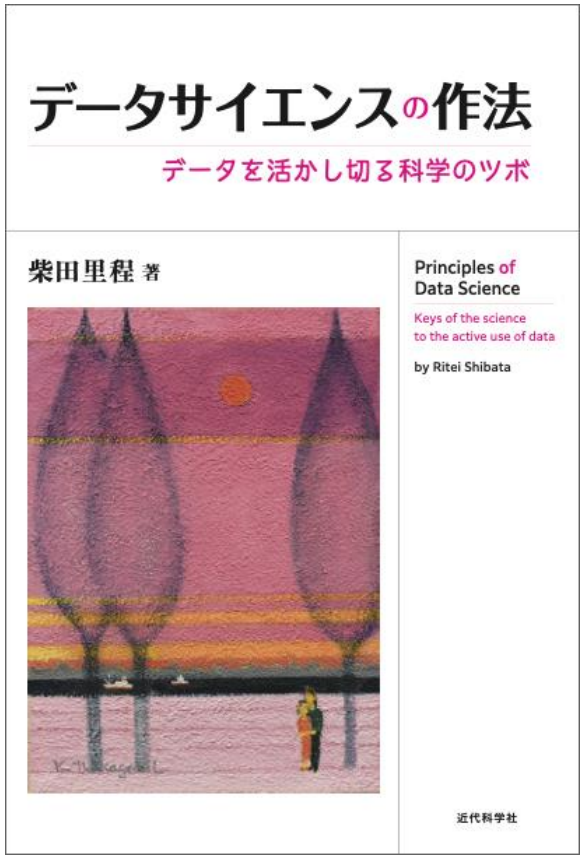
データサイエンスコンソーシアム，慶應義塾大学　柴田里程

データサイエンス研究

- 基本ソフトウエアの必要性
  - 基本的な研究環境として：知見の蓄積
  - 研究スタイルの変化にあった環境
  - 研究者間のコミュニケーションの道具として
  - 研究者と実践者の間のコミュニケーションの道具として
- 公開ソフトウエア
  - R (CUI) https://cran.r-project.org
  - TRAD（TextilePlot, R and DandD) (GUI) https://datascience.jp/TRAD.html

Data Portal : TRAD

- データの入口
  - データの概略を視覚的に把握
    - データと遊ぶ
    - 楽しむ，夢想する
    - データの攻略策を練りあげる
- TRADはデータの入口でコンシェルジュの役割も果たす
  研究者の作りあげた本格的な公開ソフトウエア
  - R はCUIで，データ全体を眺めるのは苦手
  - TRADはデータ全体をそのまま眺められる
  - EVコード署名，Apple公証済みされた安全なソフトウエア
  - Easy to use, User Friendly
  - Java で書かれた Multi-platform software
  - RoboHelp による本格的なヘルプ
  - 多言語対応

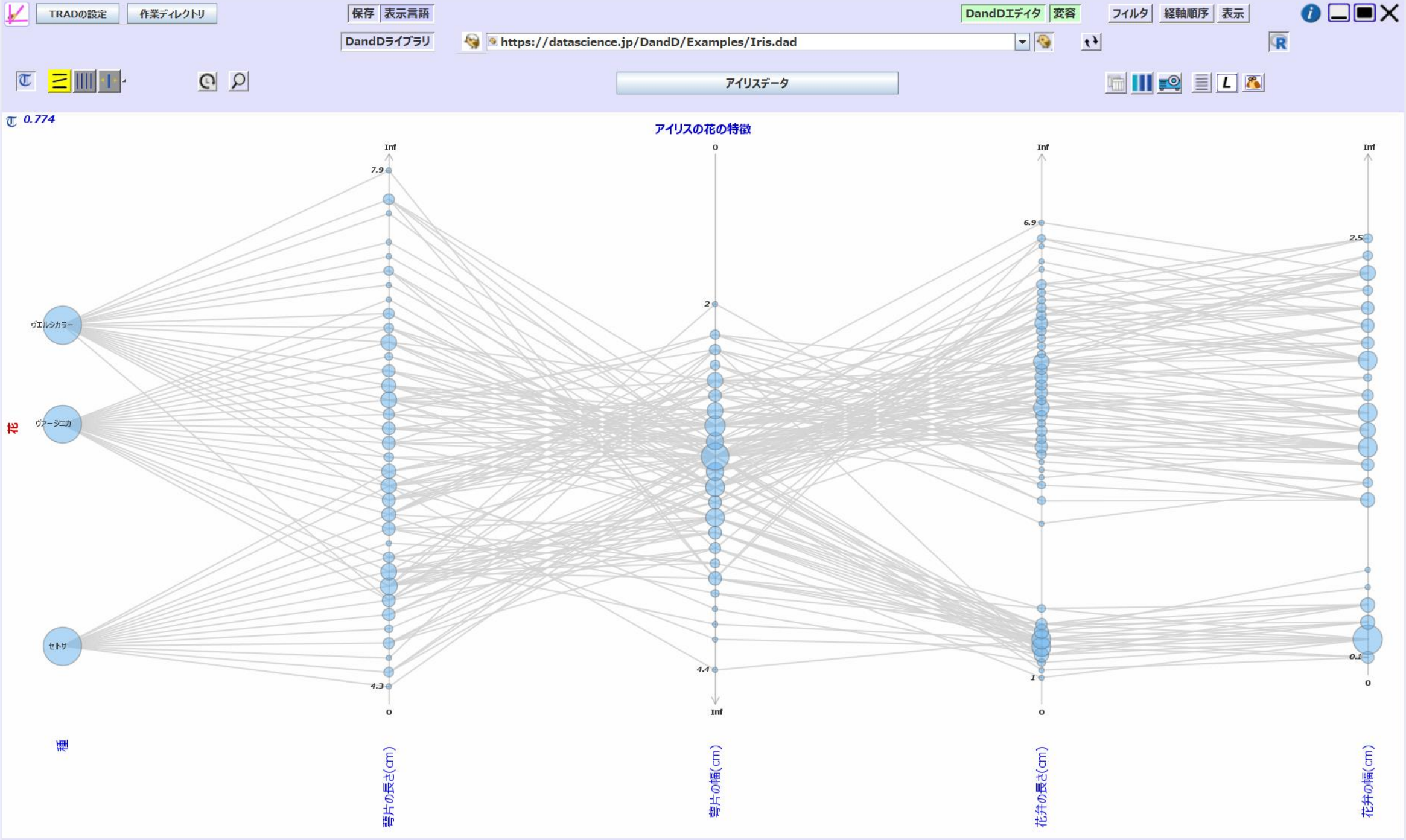


TRADの設定 作業ディレクトリ 保存 表示言語 DandDエディタ 変容 フィルタ 経軸順序 表示

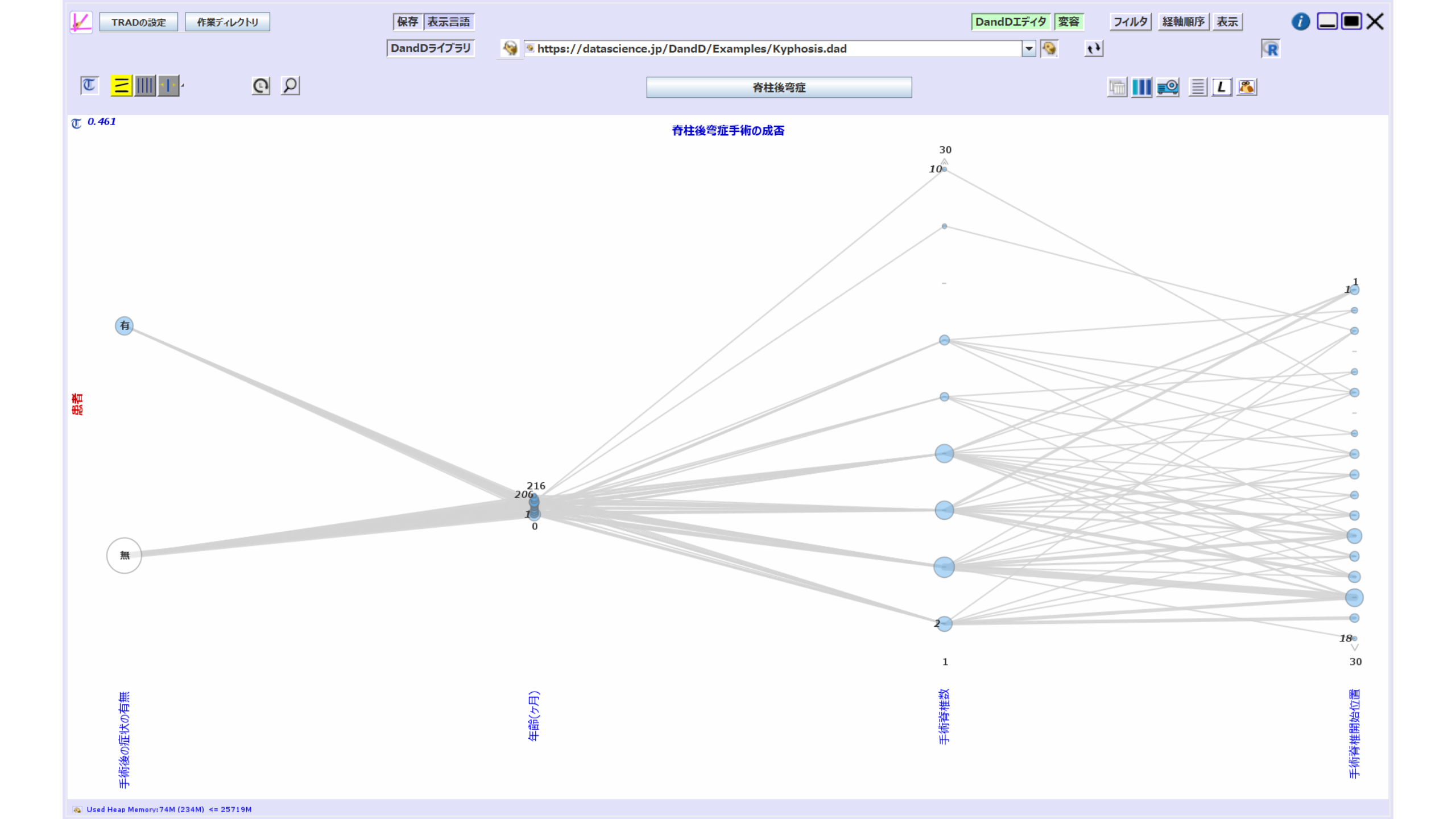
DandDライブラリ

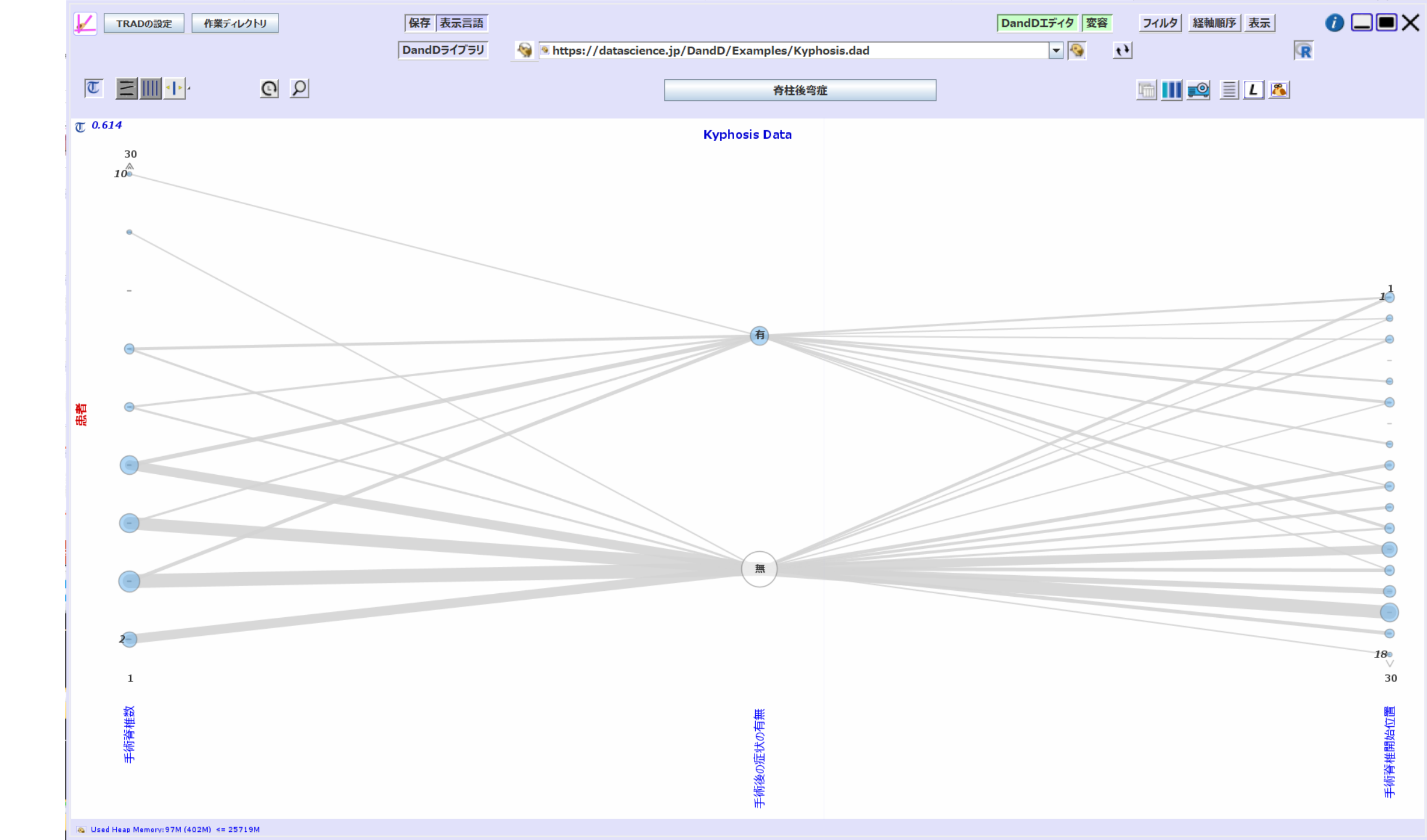
なにも読み込まれていません

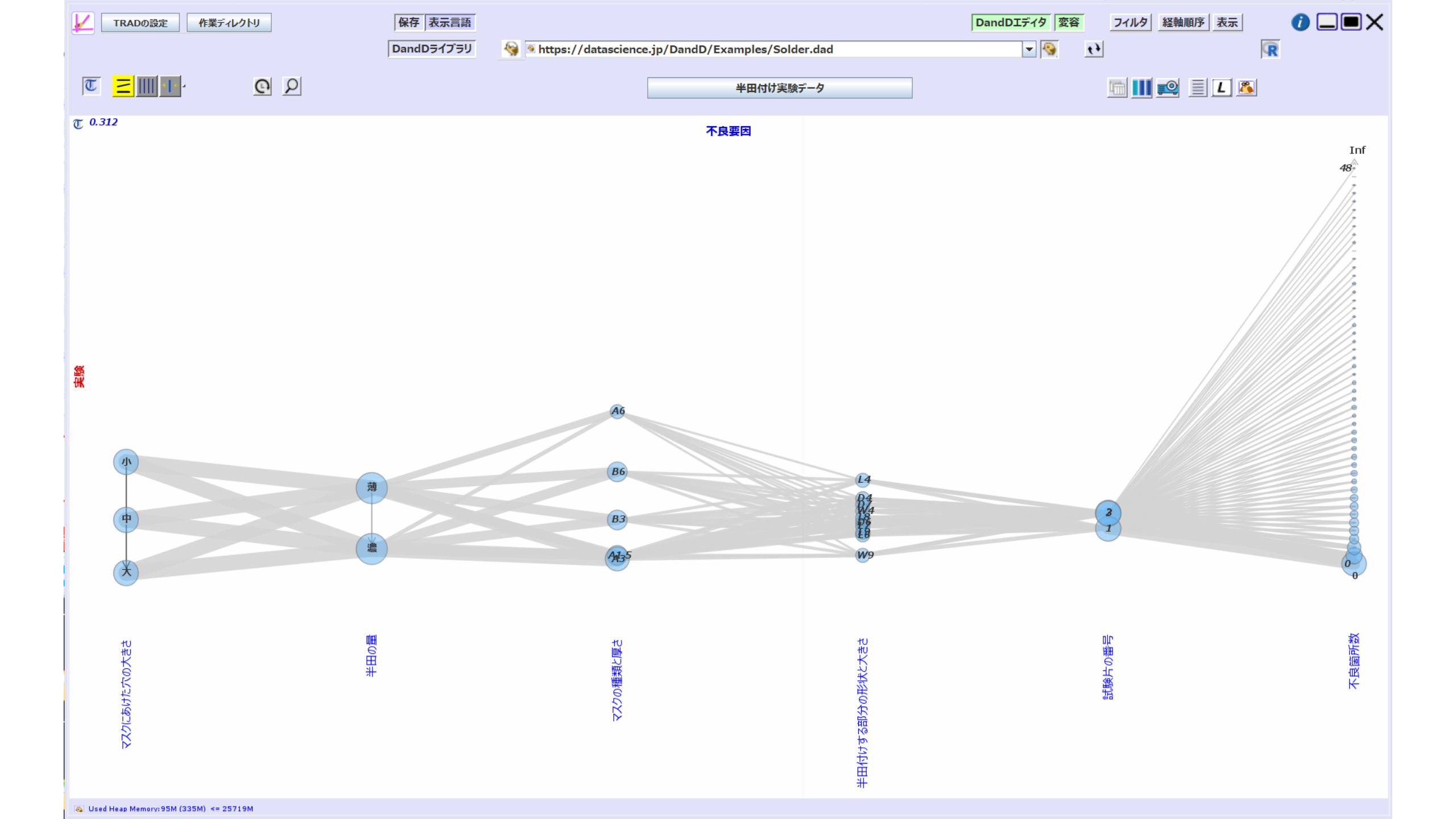
Used Heap Memory:29M (117M) <= 25719M

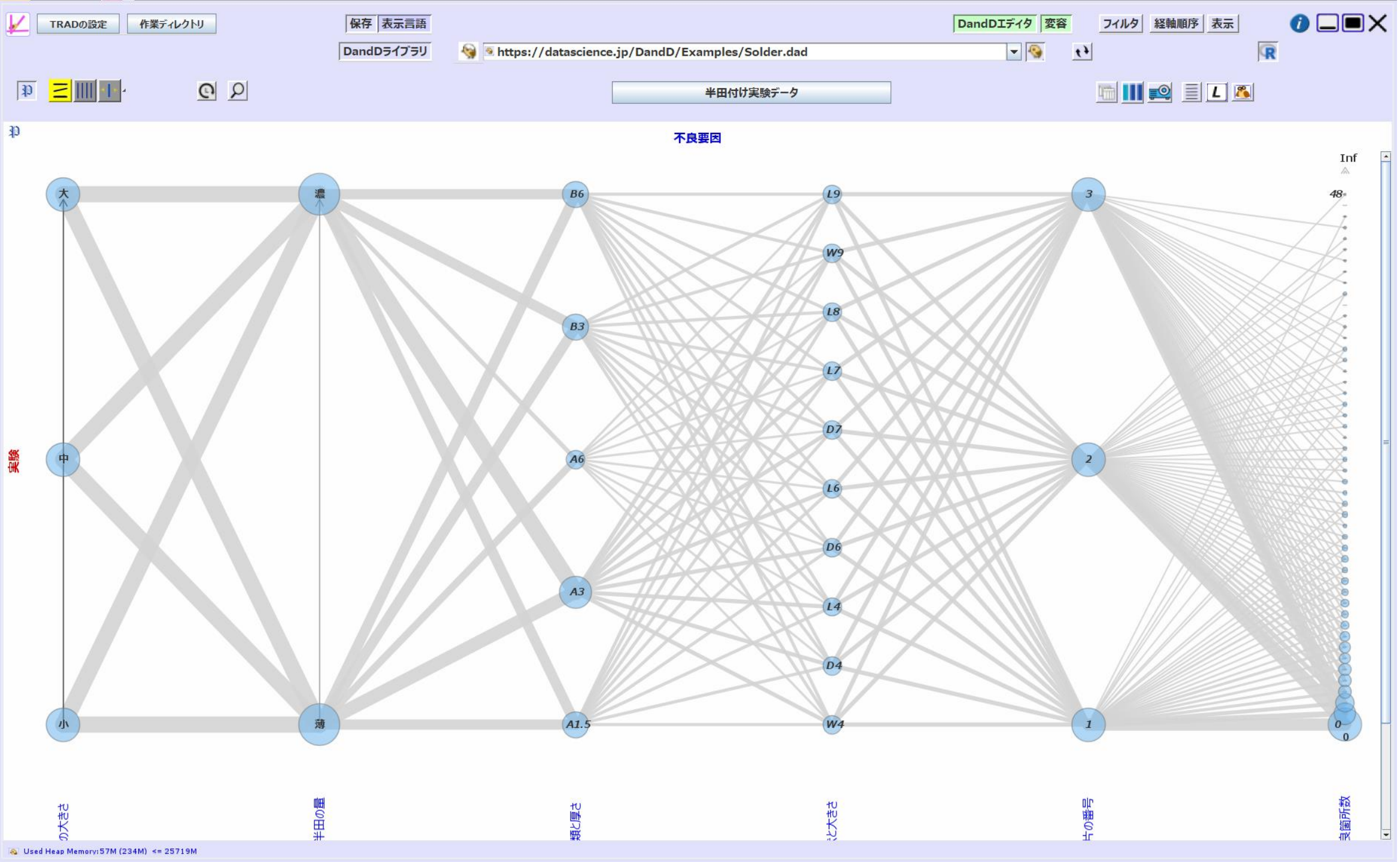

TRADの設定
作業ディレクトリ
保存
表示言語
DandDライブラリ
https://datascience.jp/DandD/Examples/Iris.dad
DandDエディタ
変容
フィルタ
経軸順序
表示
アイリスデータ
0.774
アイリスの花の特徴
花
ヴエルシカラー
ヴァージニカ
セトサ
Inf
7.9
4.3
0
0
2
4.4
Inf
Inf
6.9
1
0
Inf
2.5
0.1
0
種
萼片の長さ(cm)
萼片の幅(cm)
花弁の長さ(cm)
花弁の幅(cm)
Used Heap Memory:72M (201M) <= 25719M

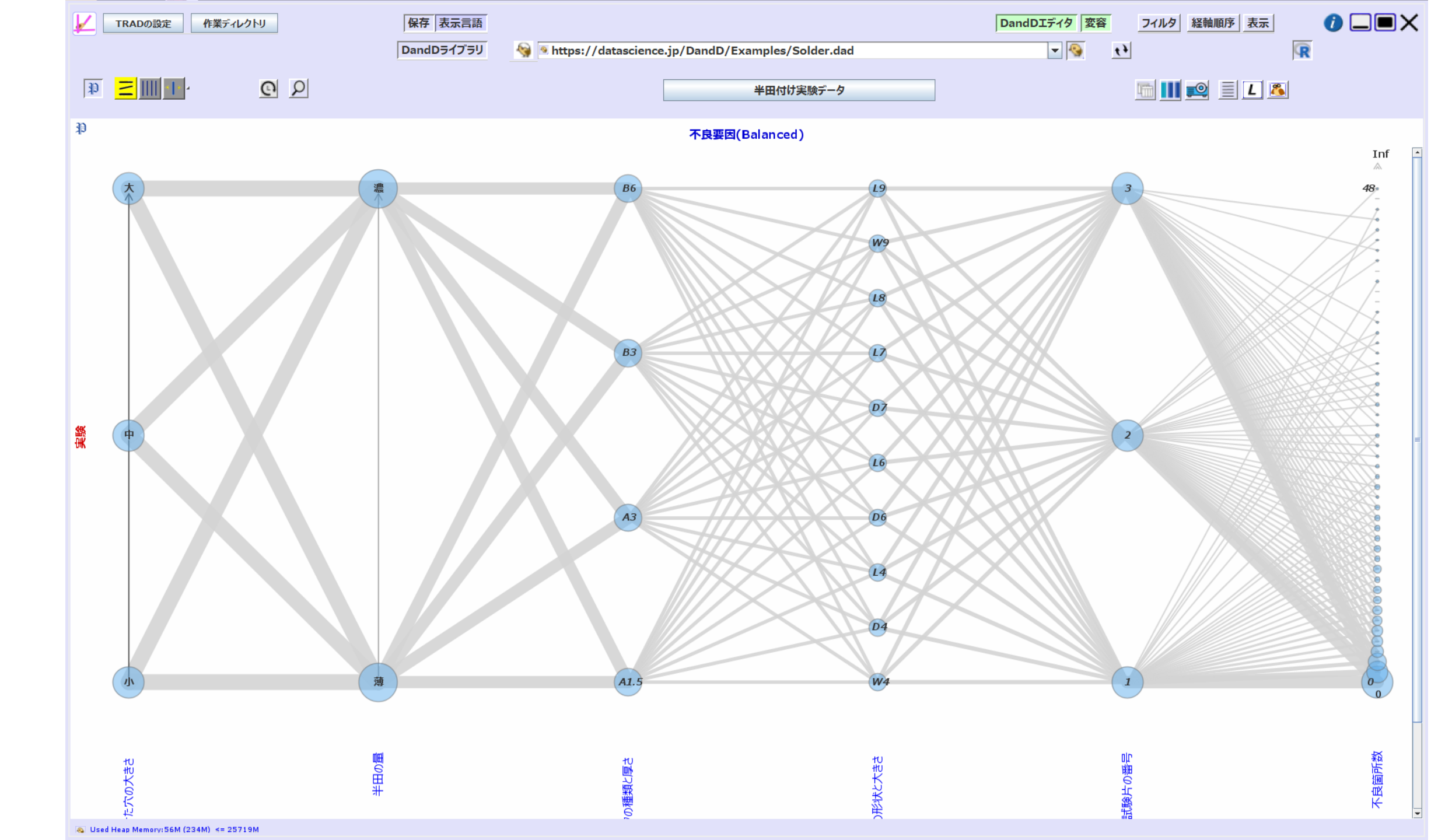

TRADの設定
作業ディレクトリ
保存
表示言語
DandDライブラリ
https://datascience.jp/DandD/Examples/Iris.dad
DandDエディタ
変容
フィルタ
経軸順序
表示
アイリスデータ
0.774
アイリスの花の特徴
Inf
7.9
4.3
0
2.5
0.1
6.9
1
ヴェルシカラー
ヴァージニカ
セトサ
2
4.4
花
萼片の長さ(cm)
花弁の幅(cm)
花弁の長さ(cm)
種
萼片の幅(cm)
Used Heap Memory:69M (201M) <= 25719M

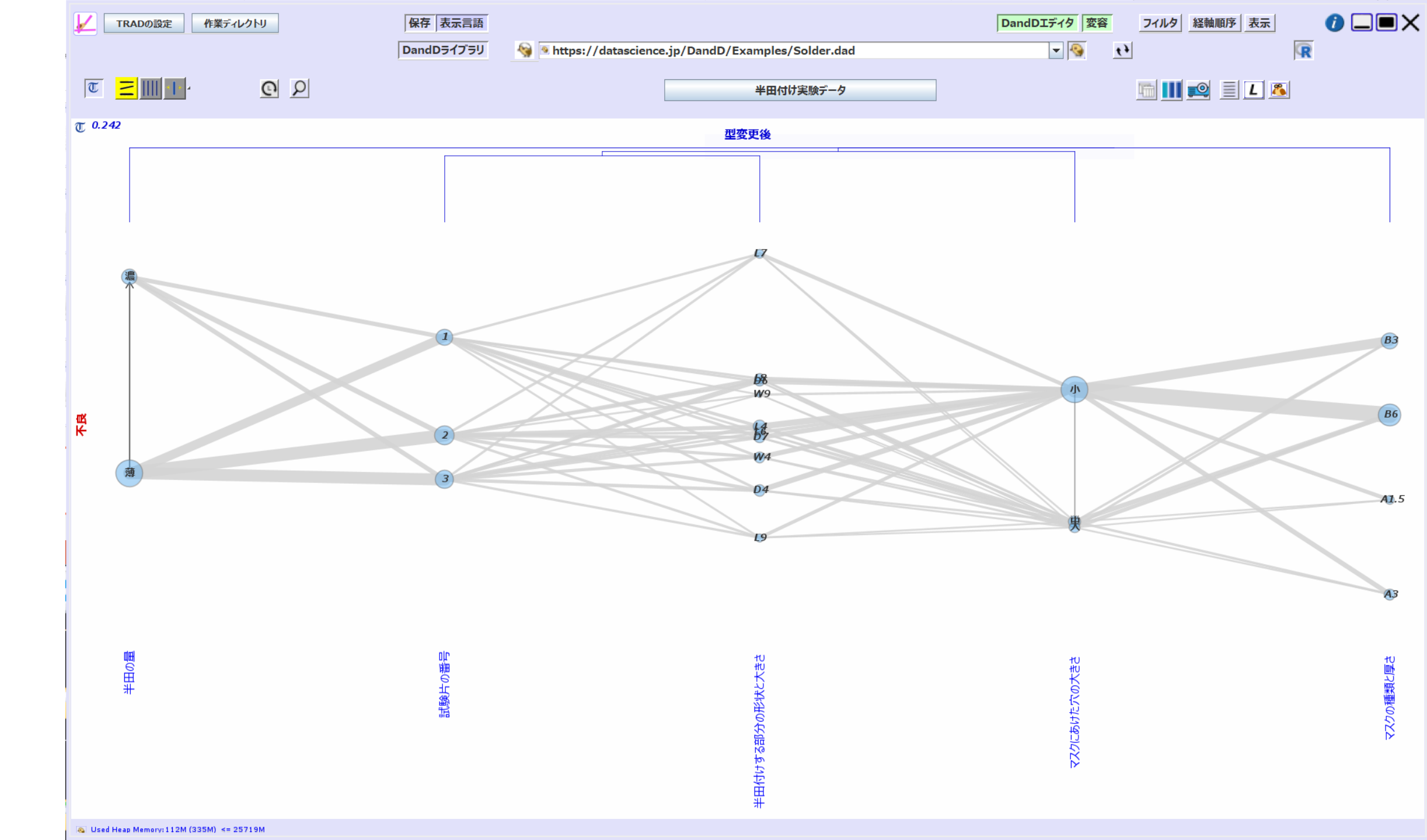

TRADの設定
作業ディレクトリ
保存
表示言語
DandDエディタ
変容
フィルタ
経軸順序
表示
DandDライブラリ
https://datascience.jp/DandD/Examples/Kyphosis.dad
脊柱後弯症
0.461
脊柱後弯症手術の成否
患者
有
無
216
206
1
0
30
10
2
1
1
1
18
30
手術後の症状の有無
年齢(ヶ月)
手術脊椎数
手術脊椎開始位置
Used Heap Memory:74M (234M) <= 25719M


TRADの設定
作業ディレクトリ
保存
表示言語
DandDライブラリ
https://datascience.jp/DandD/Examples/Kyphosis.dad
DandDエディタ
変容
フィルタ
経軸順序
表示
脊柱後弯症
0.614
Kyphosis Data
30
10
患者
1
2
有
無
1
1
18
30
手術脊椎数
手術後の症状の有無
手術脊椎開始位置
Used Heap Memory:97M (402M) <= 25719M


TRADの設定
作業ディレクトリ
保存
表示言語
DandDライブラリ
https://datascience.jp/DandD/Examples/Solder.dad
DandDエディタ
変容
フィルタ
経軸順序
表示
半田付け実験データ
0.312
不良要因
実験
Inf
48
小
中
大
薄
濃
A6
B6
B3
L4
W9
3
1
0
0
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
Used Heap Memory:95M (335M) <= 25719M


TRADの設定
作業ディレクトリ
保存
表示言語
DandDライブラリ
https://datascience.jp/DandD/Examples/Solder.dad
DandDエディタ
変容
フィルタ
経軸順序
表示
半田付け実験データ
不良要因
実験
大
中
小
濃
薄
B6
B3
A6
A3
A1.5
L9
W9
L8
L7
D7
L6
D6
L4
D4
W4
3
2
1
Inf
48
0
0
の大きさ
半田の量
類と厚さ
と大きさ
片の番号
箇所数
Used Heap Memory:57M (234M) <= 25719M


TRADの設定
作業ディレクトリ
保存
表示言語
DandDライブラリ
https://datascience.jp/DandD/Examples/Solder.dad
DandDエディタ
変容
フィルタ
経軸順序
表示
半田付け実験データ
不良要因(Balanced)
実験
大
中
小
濃
薄
B6
B3
A3
A1.5
L9
W9
L8
L7
D7
L6
D6
L4
D4
W4
3
2
1
Inf
48
0
0
半田の量
不良箇所数
Used Heap Memory:56M (234M) <= 25719M


TRADの設定
作業ディレクトリ
保存
表示言語
DandDライブラリ
https://datascience.jp/DandD/Examples/Solder.dad
DandDエディタ
変容
フィルタ
経軸順序
表示
半田付け実験データ
0.242
型変更後
不良
濃
薄
1
2
3
L7
D8
W9
L4
D7
W4
D4
L9
小
中
大
B3
B6
A1.5
A3
半田の量
試験片の番号
半田付けする部分の形状と大きさ
マスクにあけた穴の大きさ
マスクの種類と厚さ
Used Heap Memory: 112M (335M) <= 25719M

TextilePlot表示

- 並行座標プロットの一種
- 座標を緯糸で結んで一記録を表している
- 各軸のローケーション，スケールを水平基準で緯糸がなるべく水平になるように定めているだけ.
- 軸（変量）のクラスタリングは，軸を結ぶ緯糸の傾きの絶対和を距離とする最小距離法（single linkage)
- 各変量の型を反映した表示

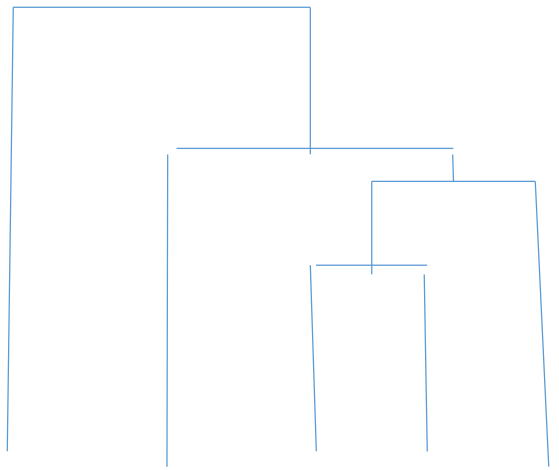
大規模データ

- 記録数大
  - Thinning, 縦スクロール
- 変量数大
  - 横スクロール
  - （変量の）クラスタリングアルゴリズム
    - 変量数を p としたとき
      - ナイーブなアルゴリズム: 計算時間$O(p^3)$, メモリ$O(p)$
      - SLINK: 計算時間$O(p^2)$, メモリ$O(p)$

SLINK アルゴリズム(Sibson, 1973)

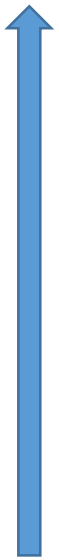
最小距離による階層的クラスタリング

例： $p = 5,\ i = 0,1,2,3,4$



$\tau(i)$　2　1　0　3　4

並べ替え

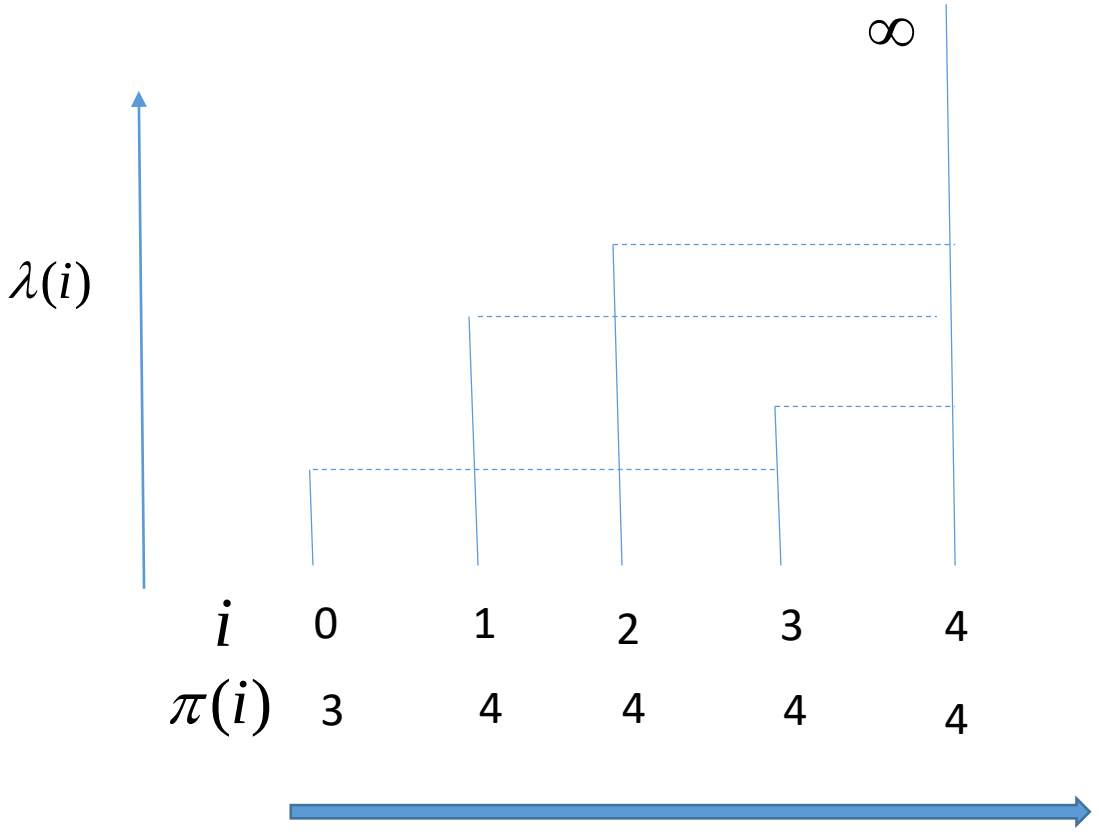


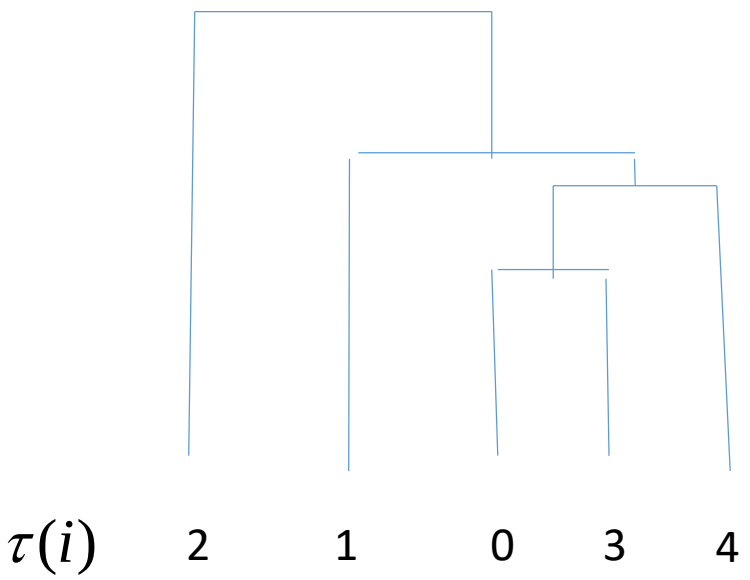
ナイーブなアルゴリズム

$\pi(i) = i$があるクラスタに加わったときの

　　そのクラスタ内の最大要素番号(クラスタ番号)

$\lambda(i)$ = その時の距離





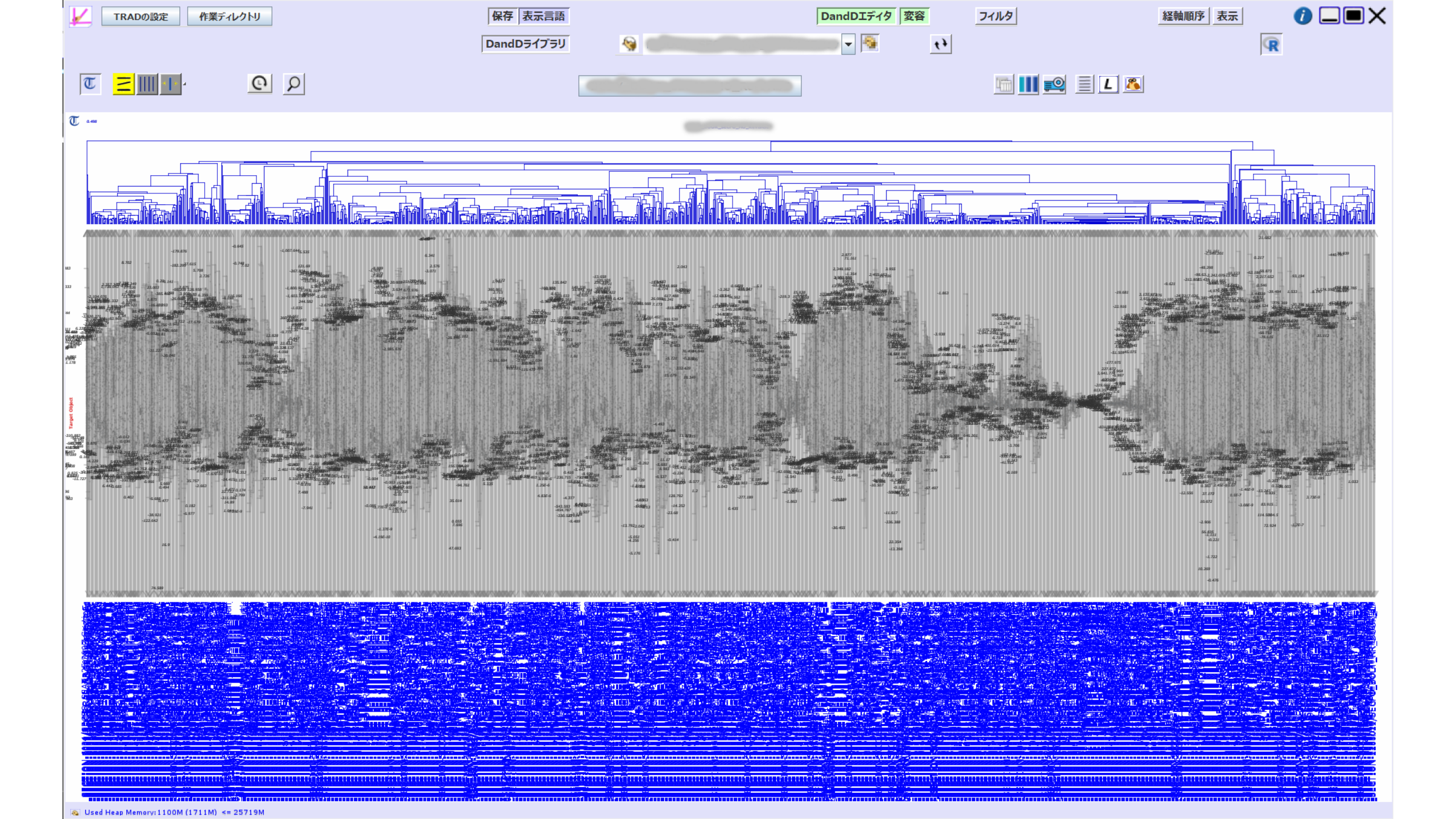
事後的に変量数が大きくなる例

$X_1, X_2, ..., X_m$：説明変量（条件），$Z_1(t), Z_2(t), ..., Z_p(t)$：時系列（計測値）

このような実験がn回行われたとき，その関係を眺める一つの方法は
各時系列をいくつかの時点$t_1, t_2, ..., t_k$での値で代表させ，$m + pk$変量

$X_1, X_2, ..., X_m$，$Z_1(t_1), Z_1(t_2), ..., Z_1(t_k)$，$Z_2(t_1), Z_2(t_2), ..., Z_2(t_k)$，...，$Z_p(t_1), Z_p(t_2), ..., Z_p(t_k)$

のn記録として眺めること．


TRADの設定
作業ディレクトリ
保存
表示言語
DandDエディタ
変容
フィルタ
経軸順序
表示
DandDライブラリ
Target Object
Used Heap Memory:1100M (1711M) <= 25719M