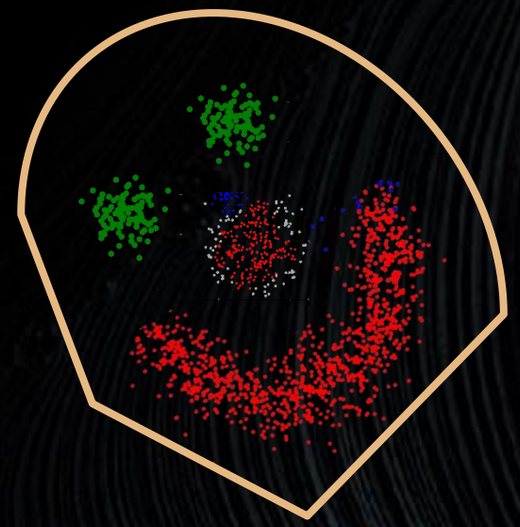
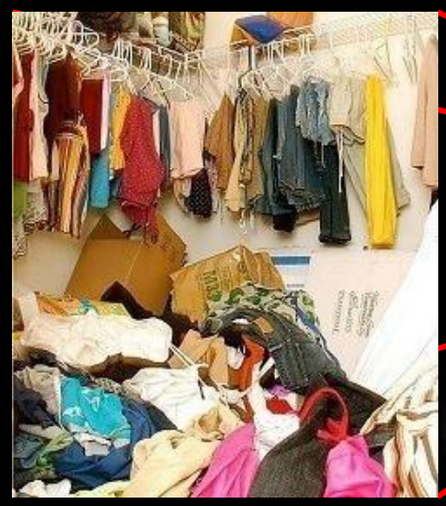
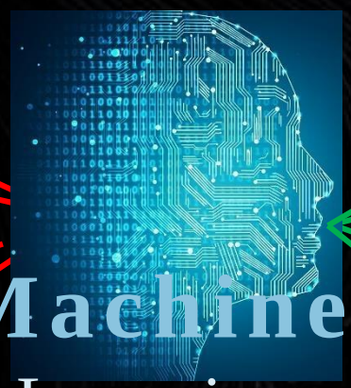
Unsupervised learning: Principal Component and Cluster Analysis

Unsupervised learning: Principal Component and Cluster Analysis







Machine
Learning

Exploratory methods: unsupervised

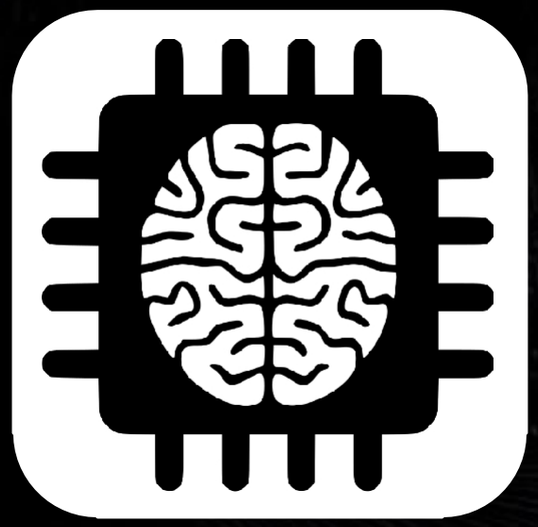
national geographic

Exploratory methods: unsupervised

ACP

Description

Summary



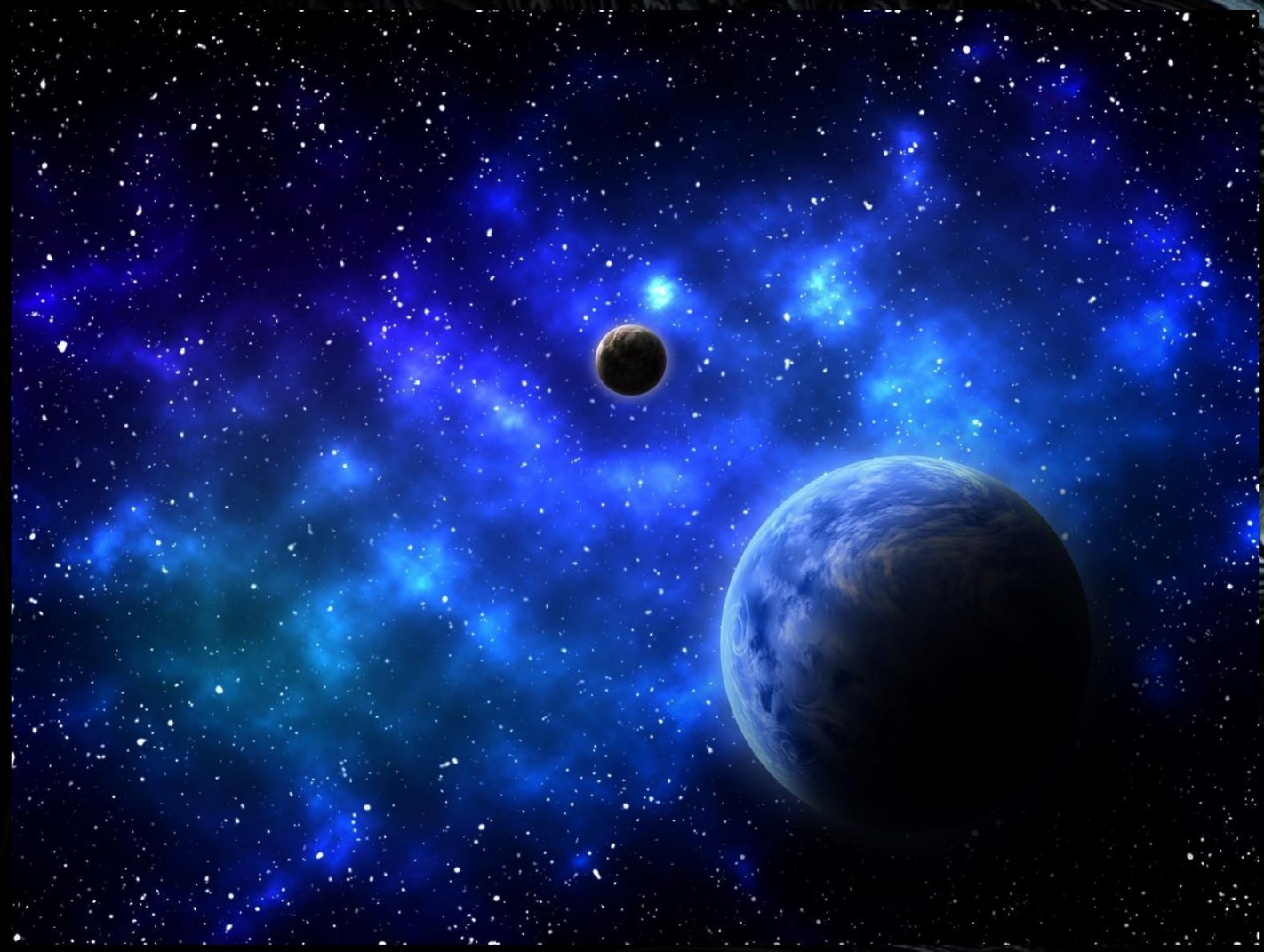
Cluster

Partitioning the dataset

How ACP works

Data space: a multidimensional space

As many dimensions as variables

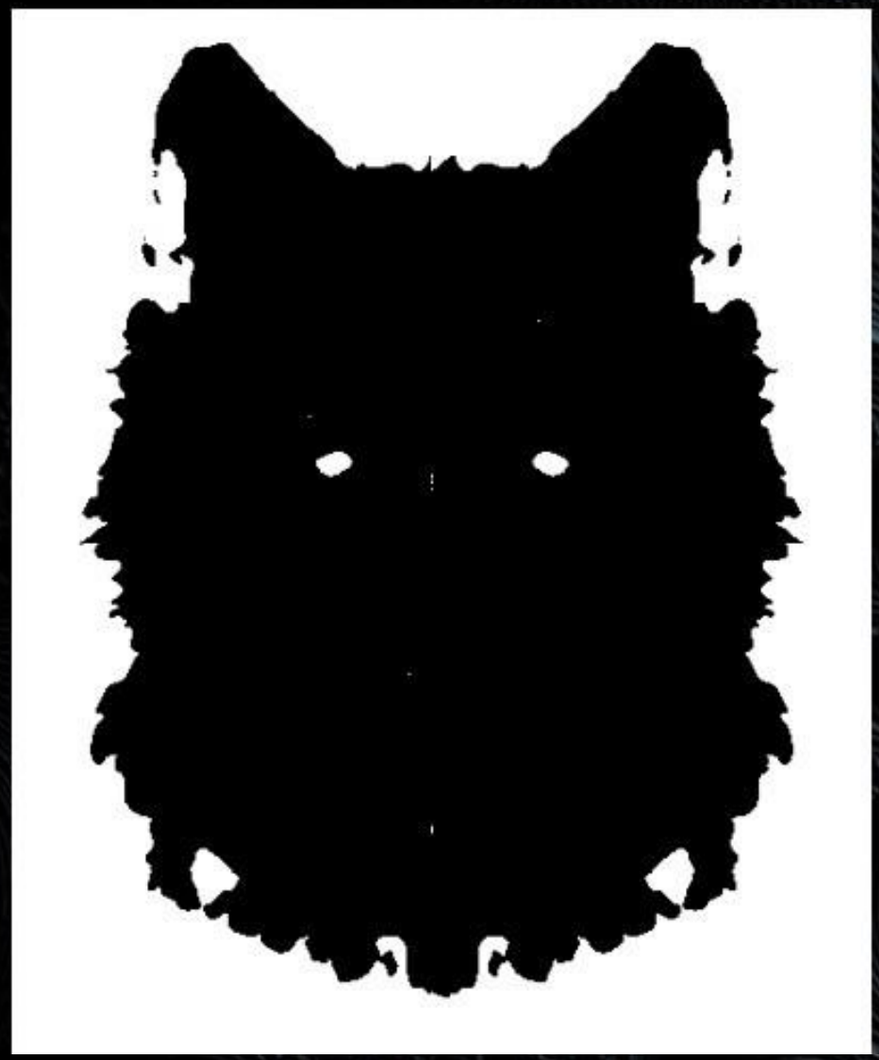


How ACP works

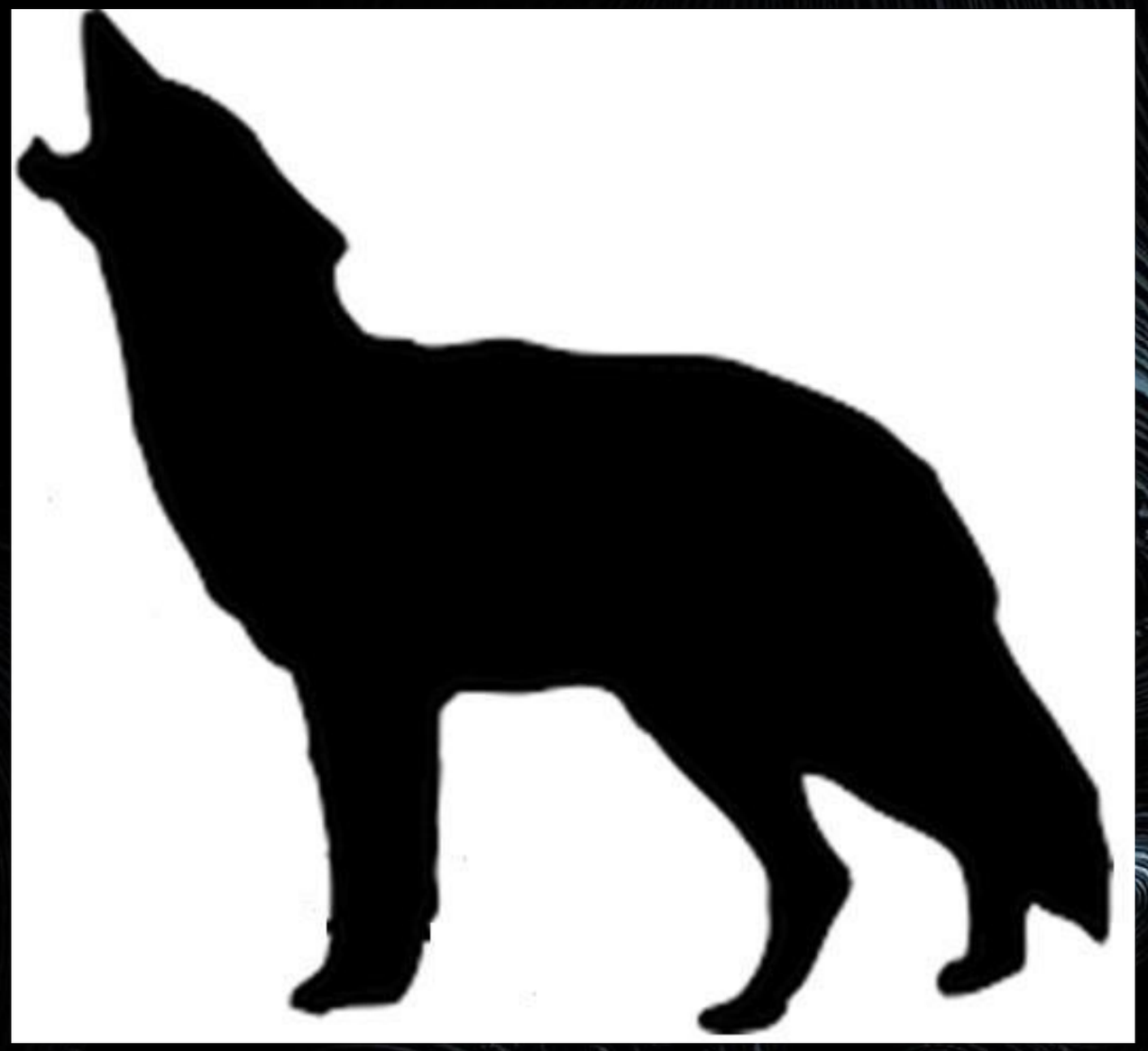
Multidimensional space

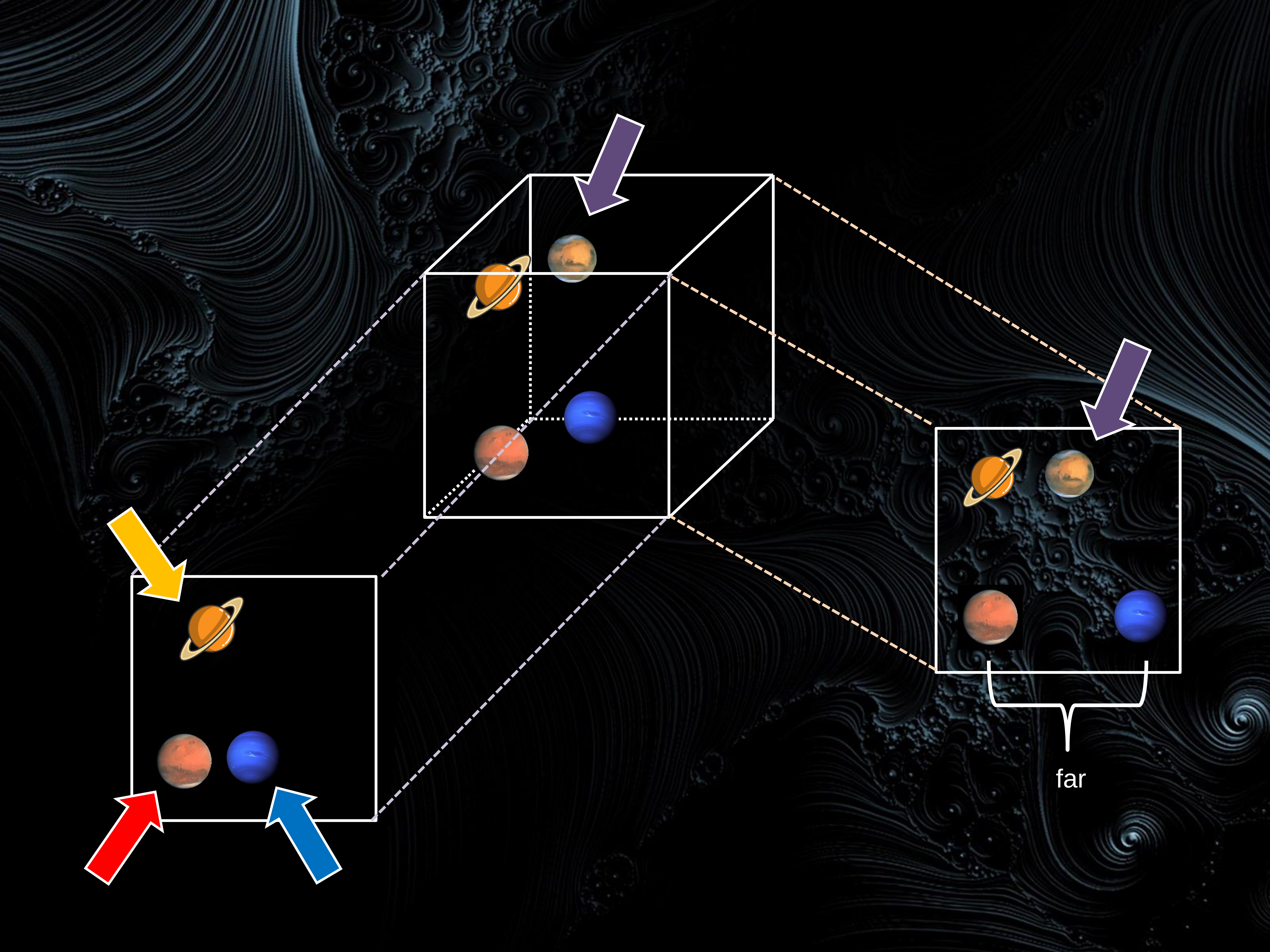
How ACP works

How ACP works



How ACP works

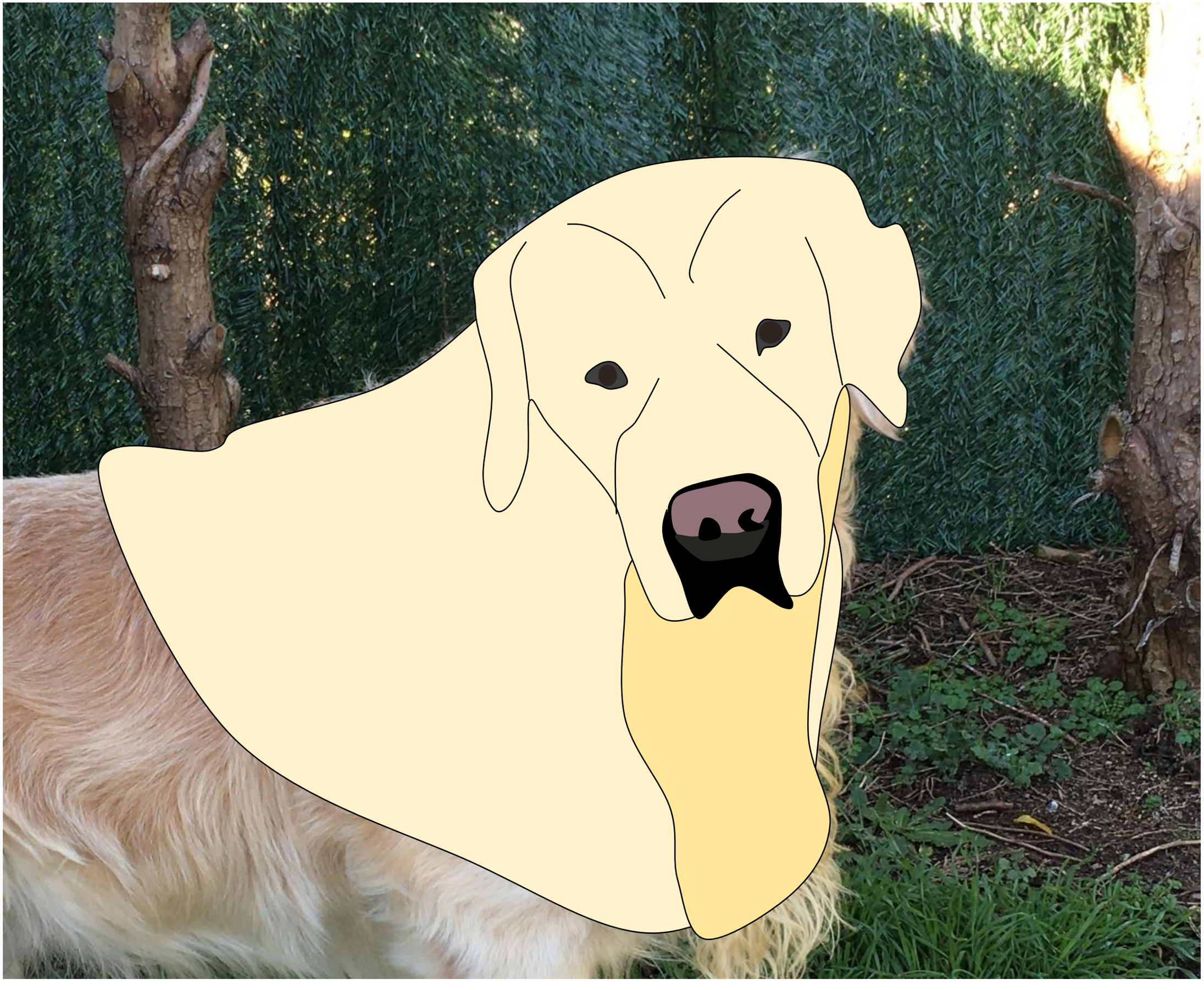


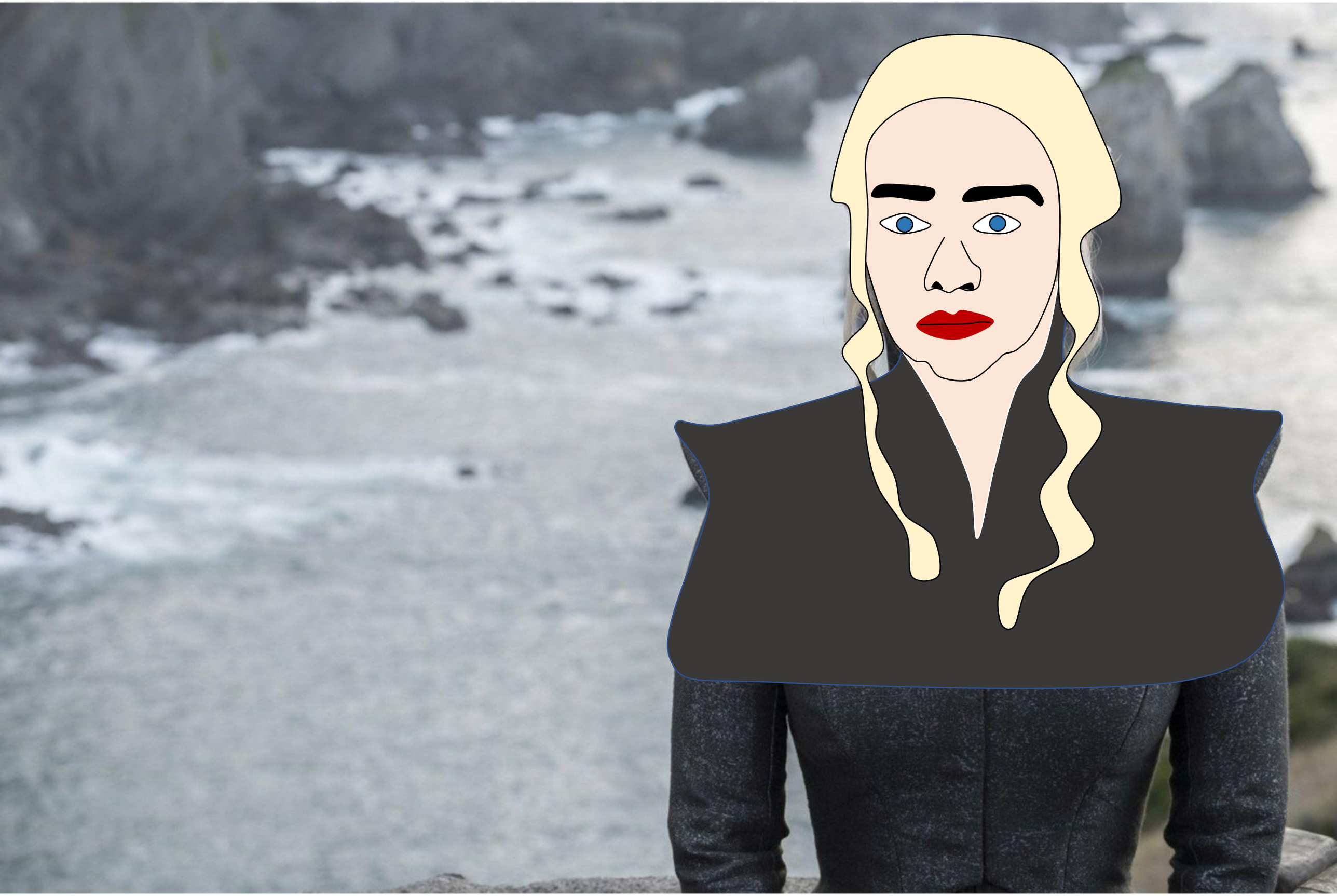

far

How ACP works

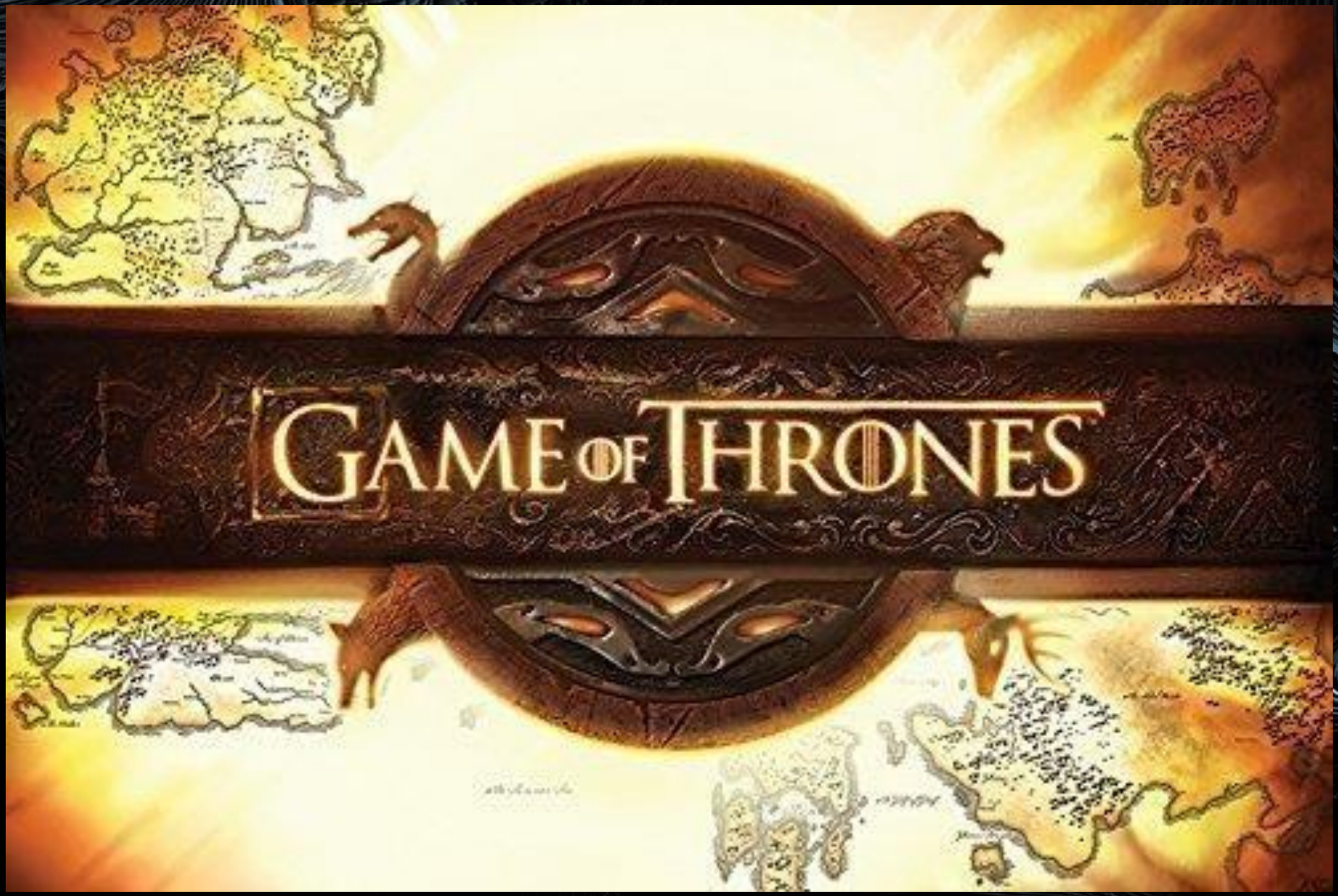
ACP = satellite that takes photos of multidimensional space and ranks them from most to least informative

Data space simplification





Dataset (Kaggle)



Objective and data set

- Character similarity criteria

PCA visualization

variables

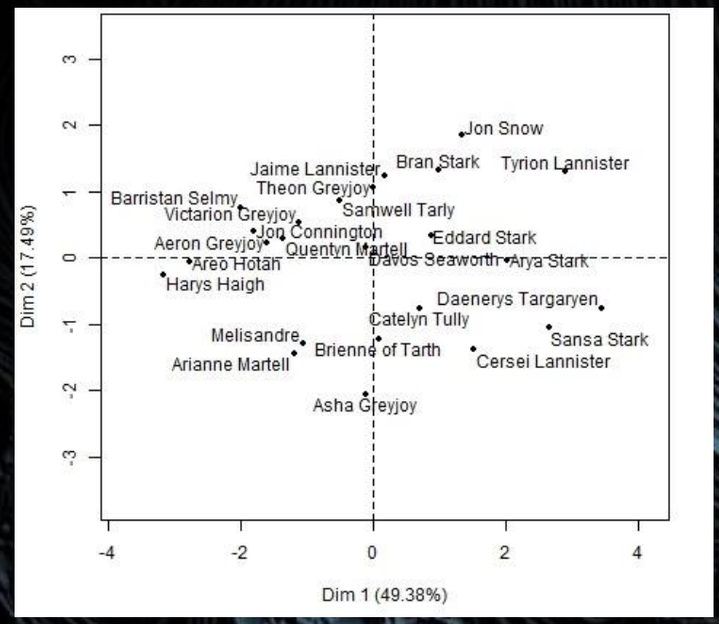
characters

	var 1	Var 2	var 3	Var 4
Ind1				
Ind 2				
Ind 3				
Ind 4				
Ind 5				

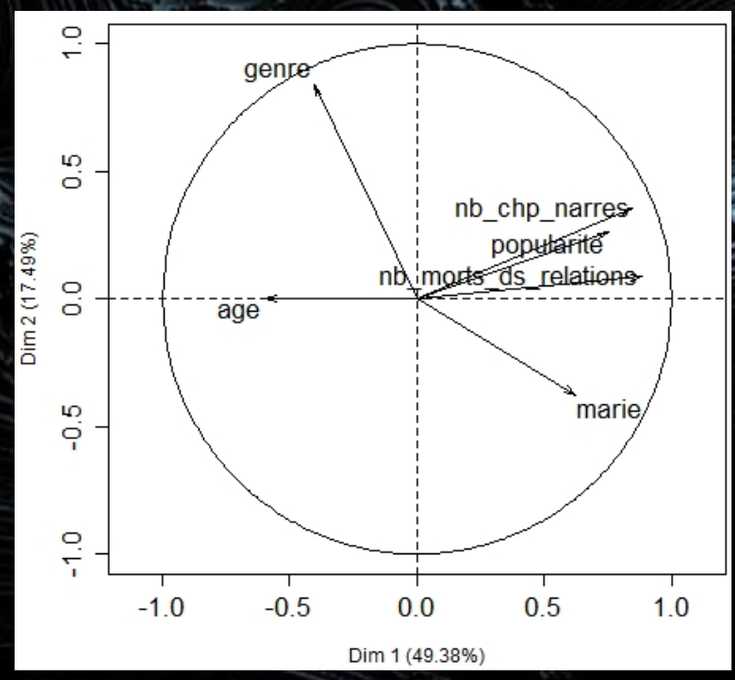
- Plot of individuals
- Plot of variables

Two types of plots

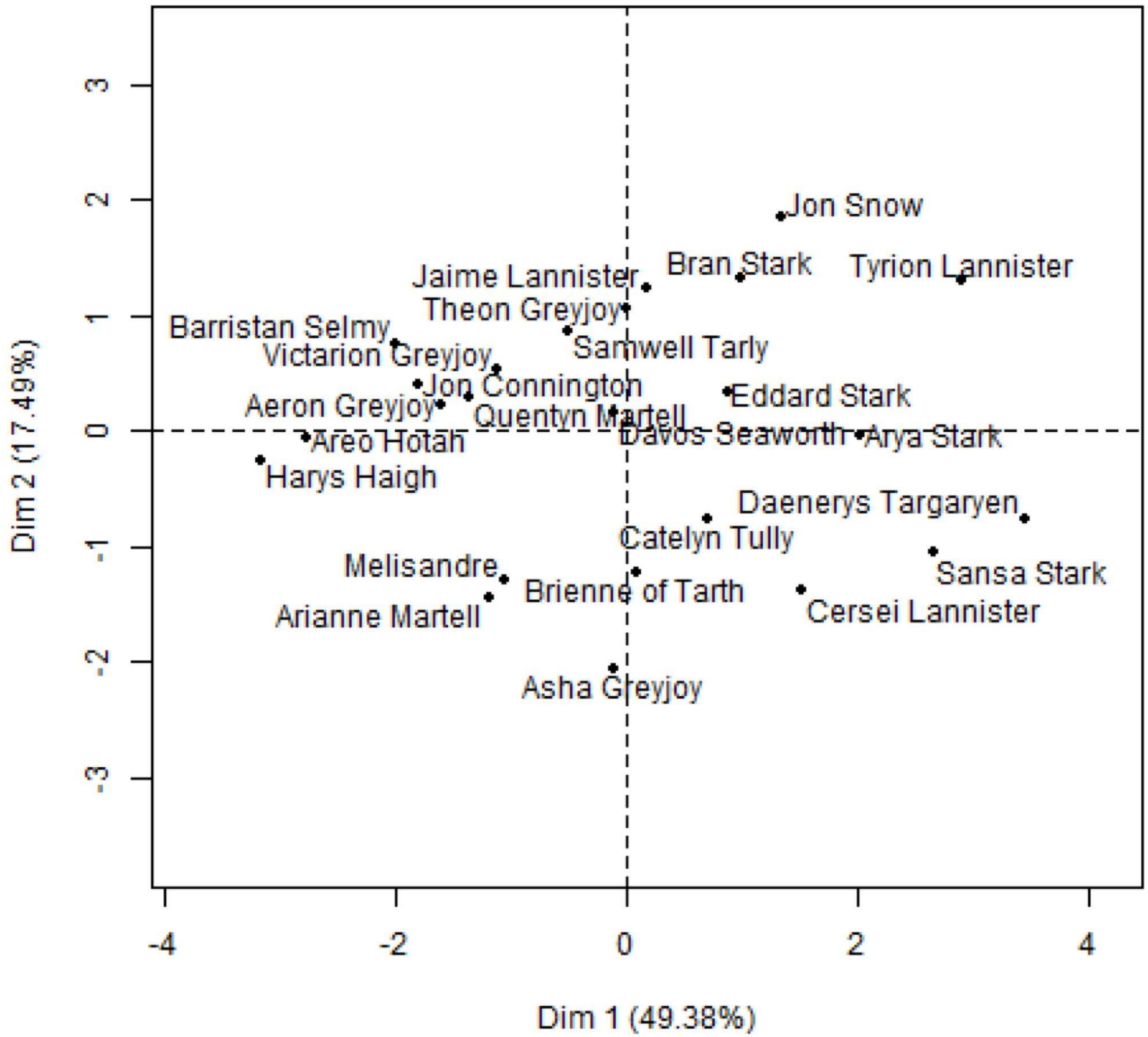
- The plot of individuals



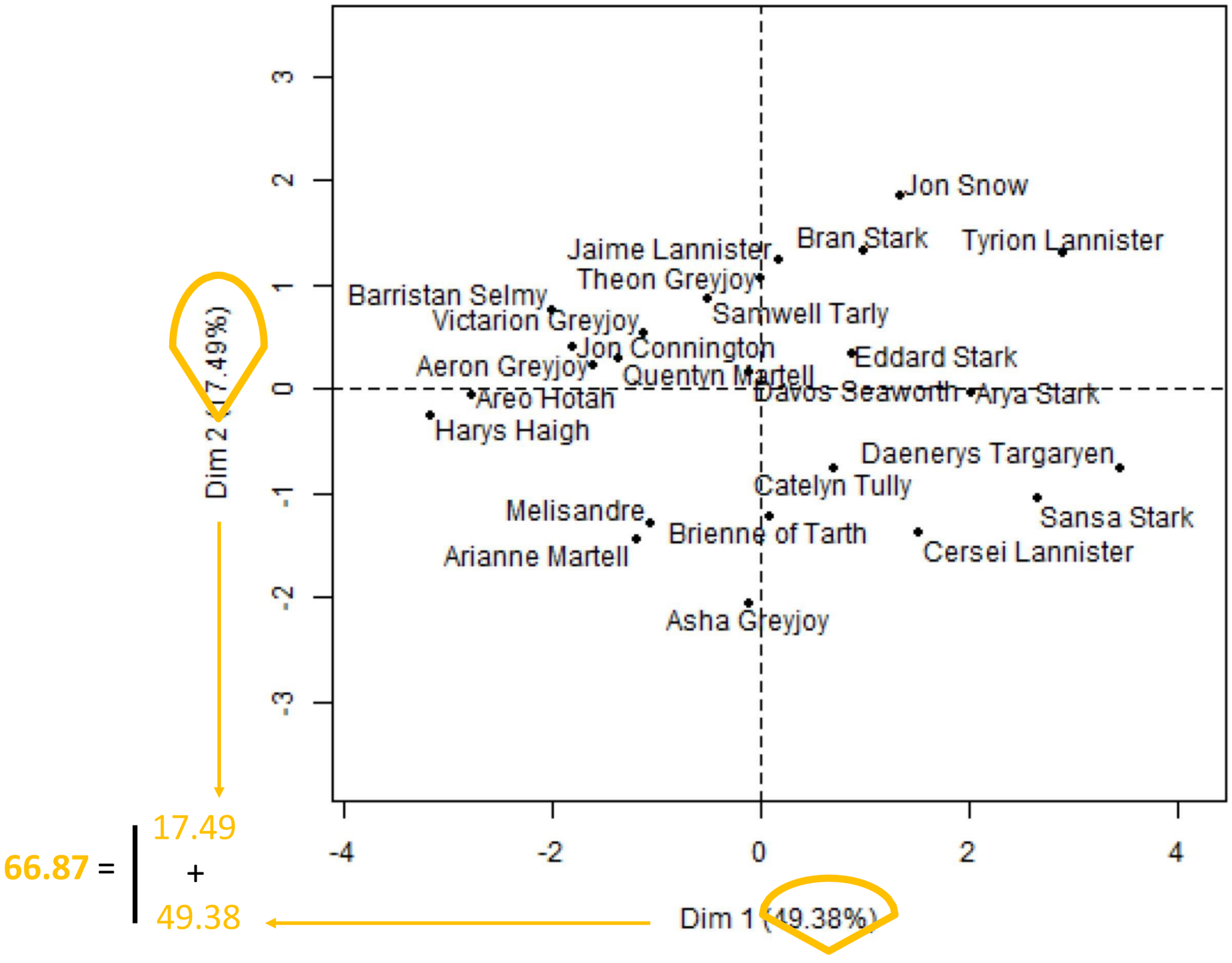
- The plot of variables = circle of correlations



The plot of individuals



The plot of individuals

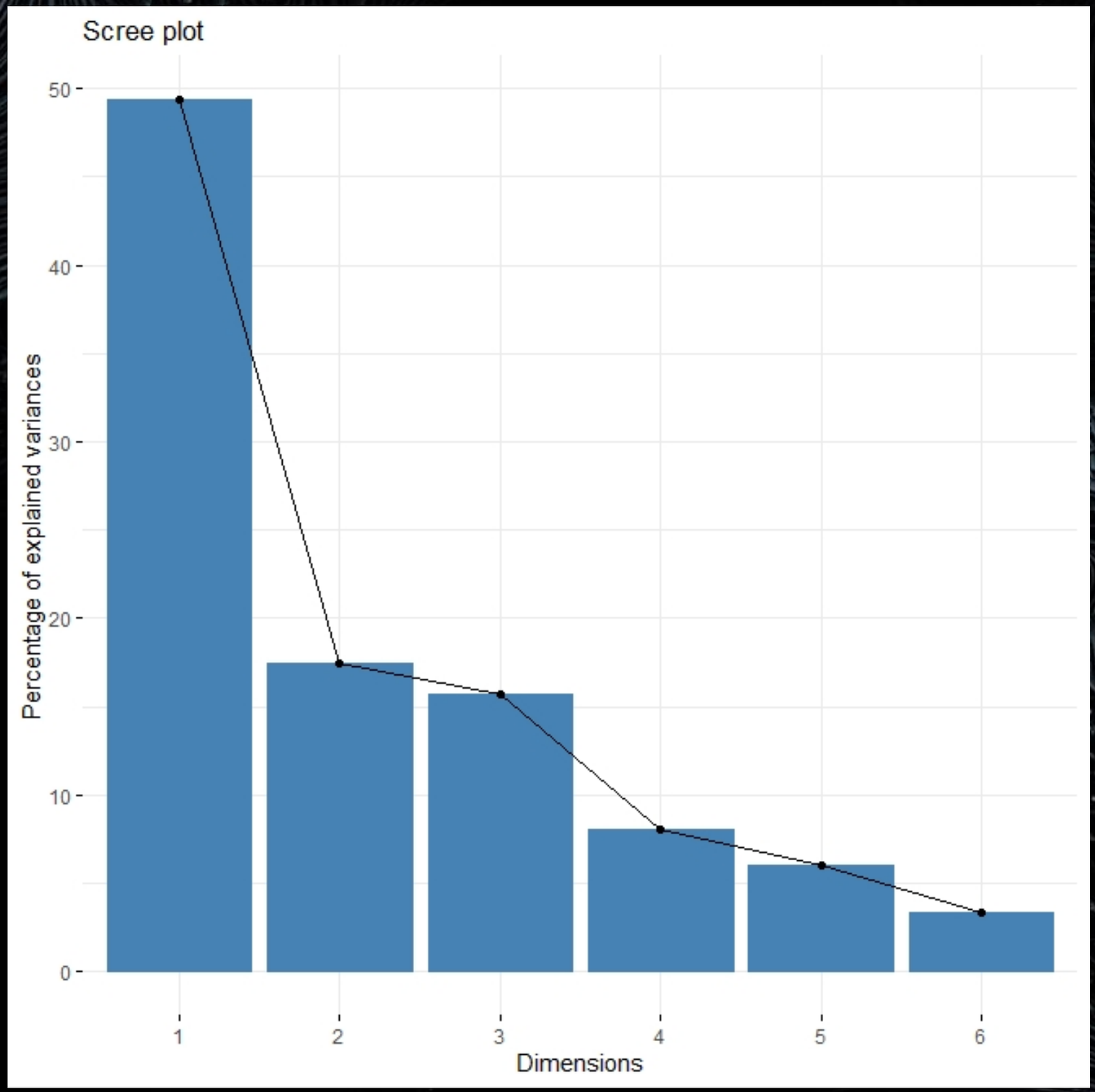


66.87 = 17.49 + 49.38

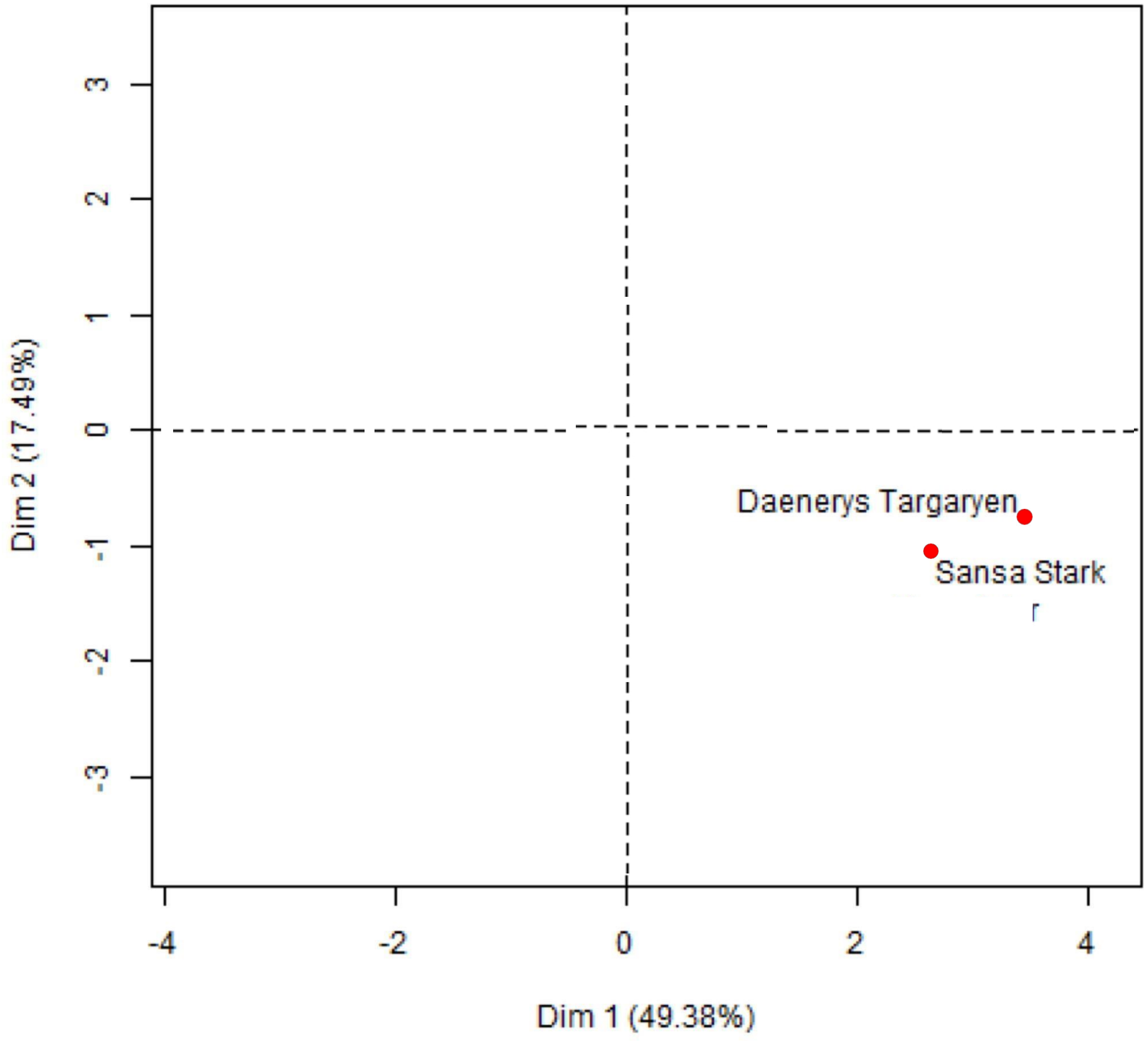
% inertia ~ % information

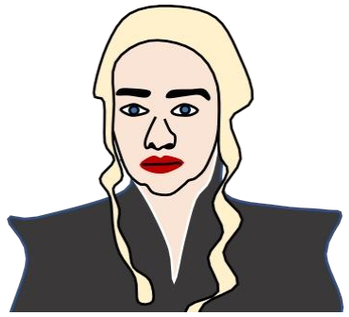
Is it a tragedy to have a low percentage of inertia?

Inertia decay

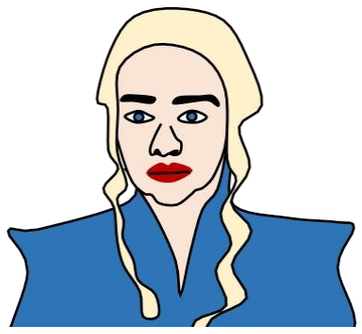


Distance is key!

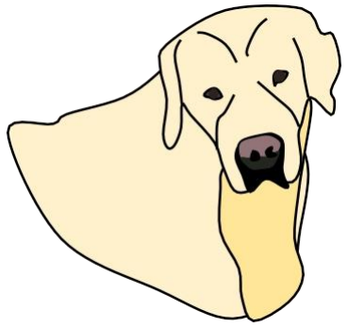

Daenerys Targaryen
Sansa Stark
Dim 2 (17.49%)
Dim 1 (49.38%)
3
2
1
0
-1
-2
-3
-4
-2
0
2
4



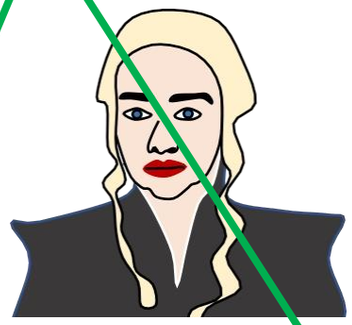
Daenerys in **Black**



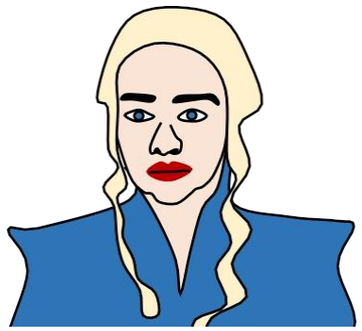
Daenerys in **Blue**



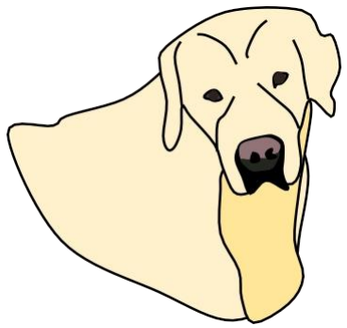
Dog



Daenerys in **Black**

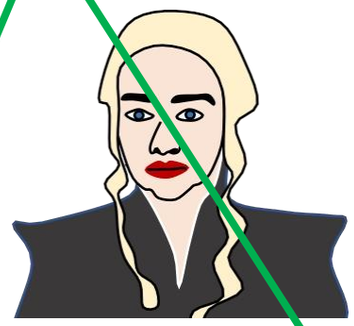


Daenerys in **Blue**

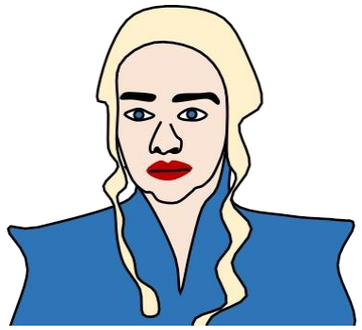


Dog

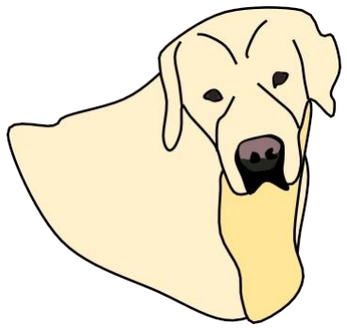
The two are very similar



Daenerys in **Black**



Daenerys in **Blue**



Dog

The two are very similar

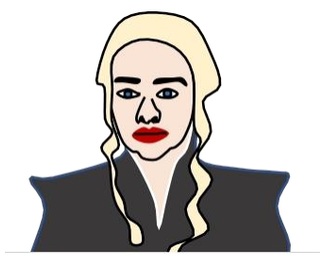
PCA

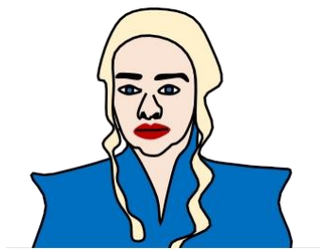
Dim 2

PCA

Dim 1

Dim 2

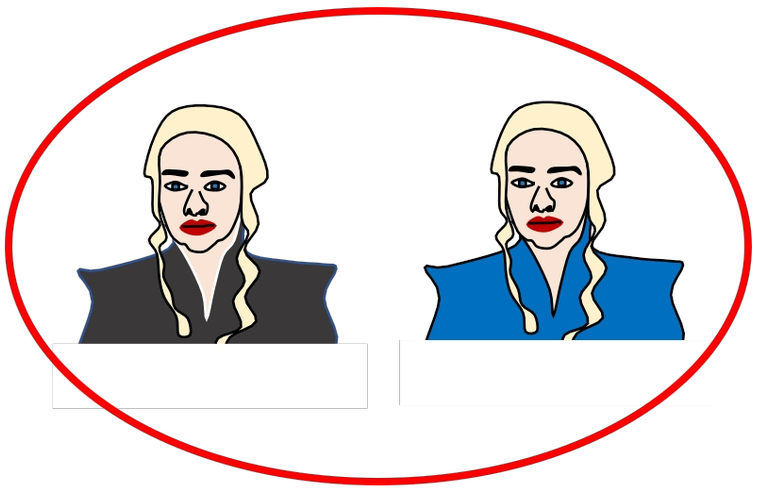




Dim 1

Dim 2

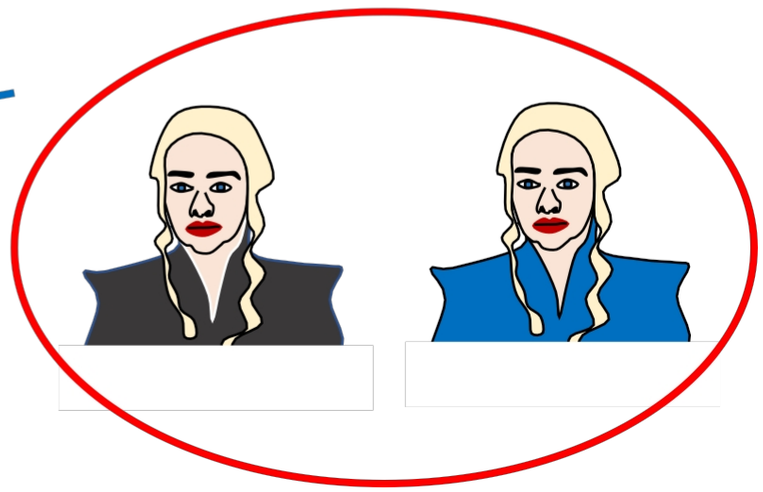
Close = Similar



Dim 1

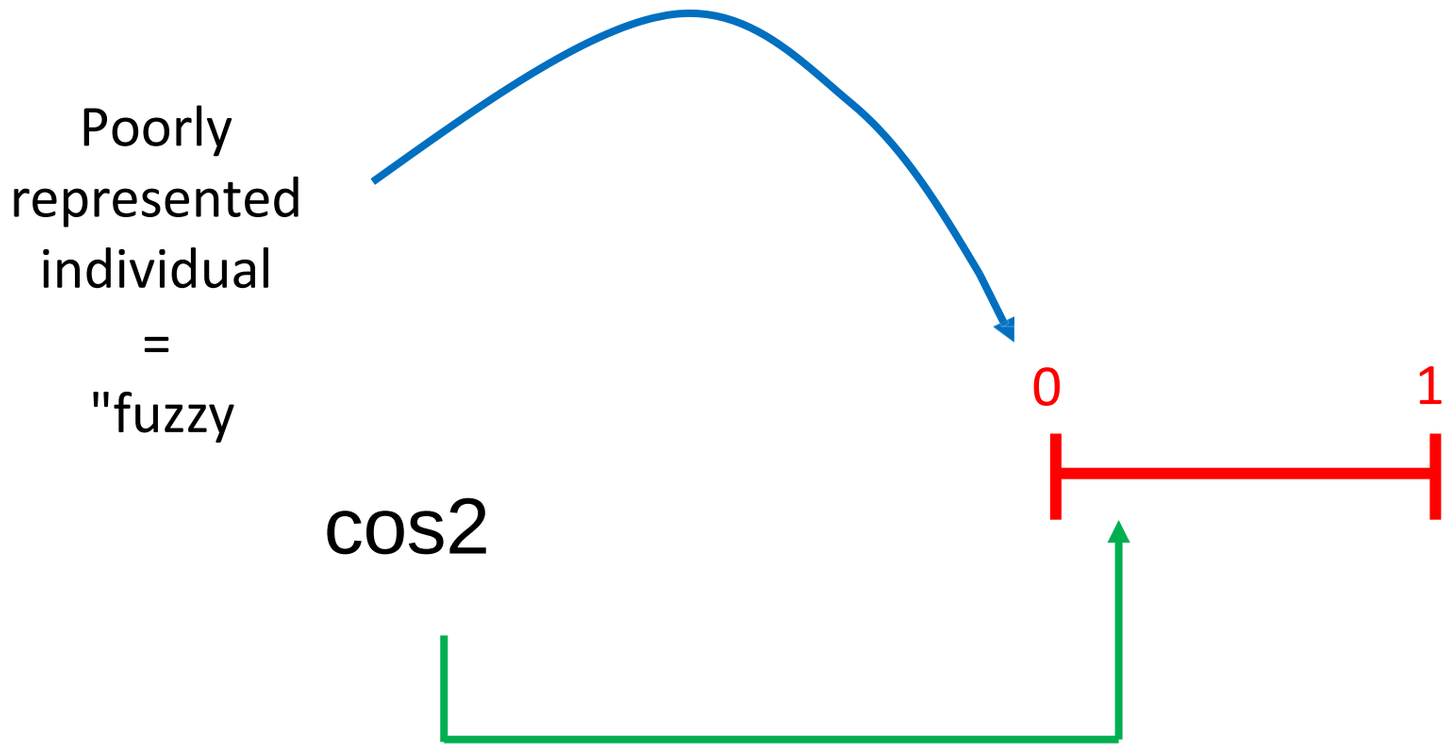
Far = different

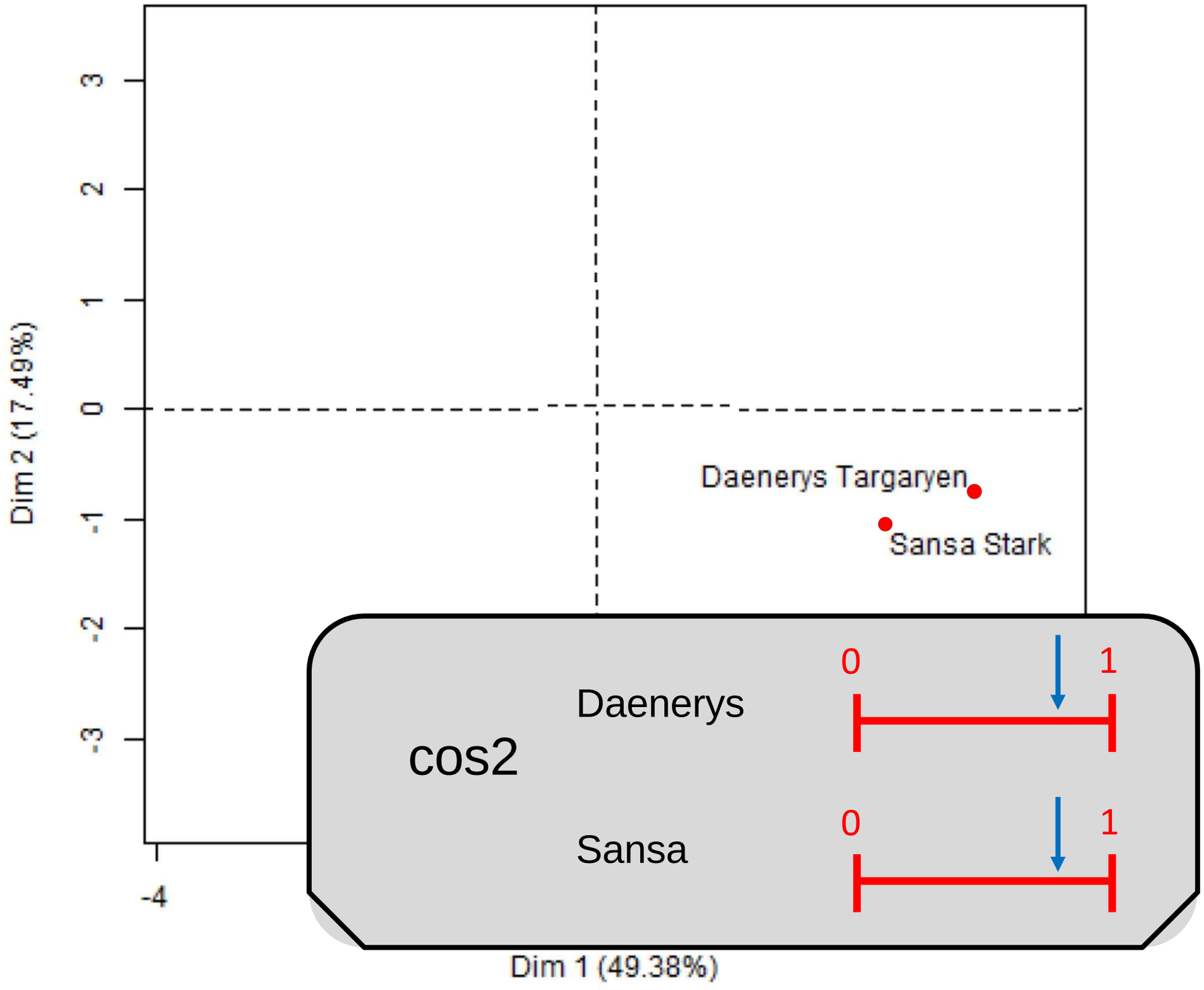
Close = Similar

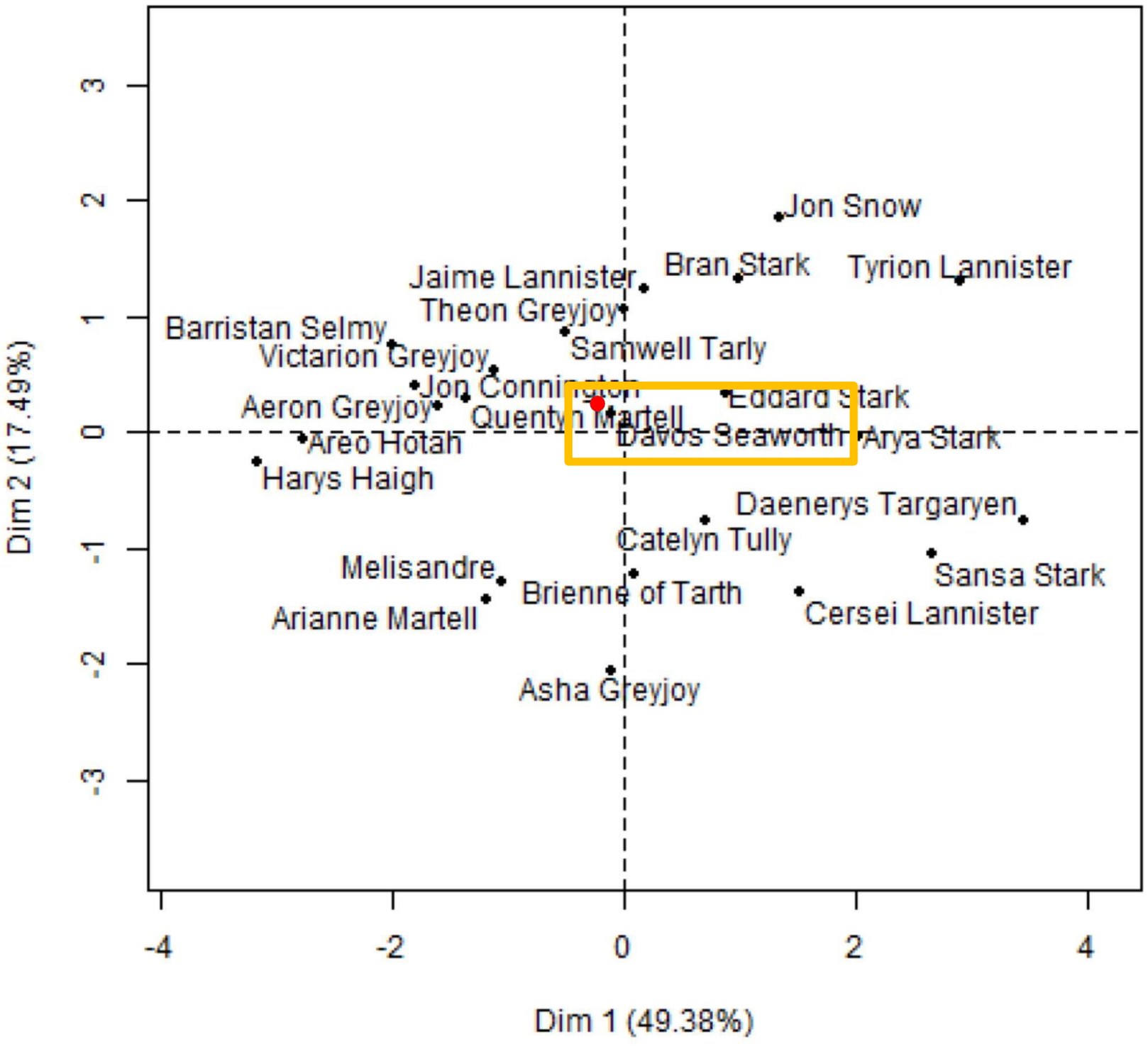


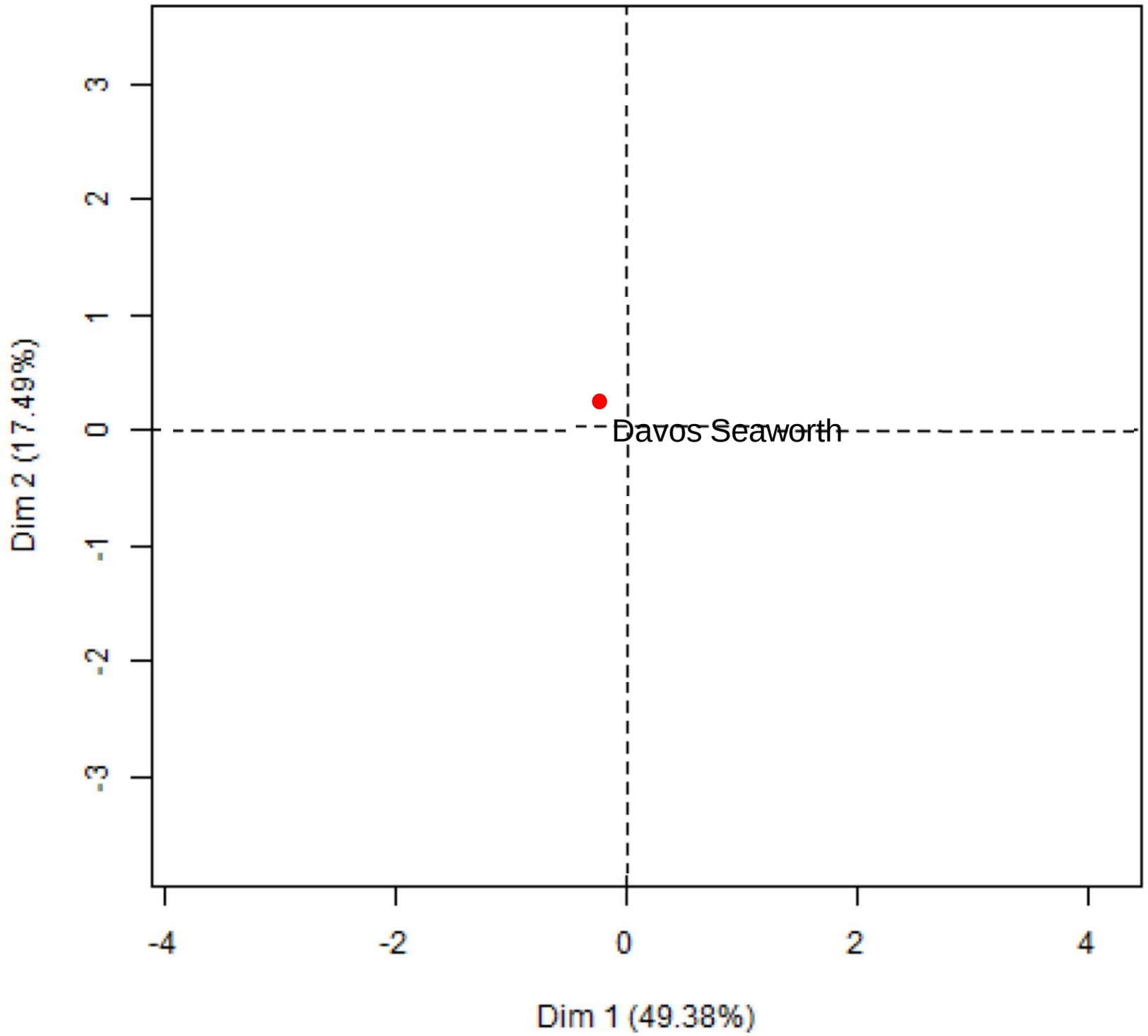
Dim 2

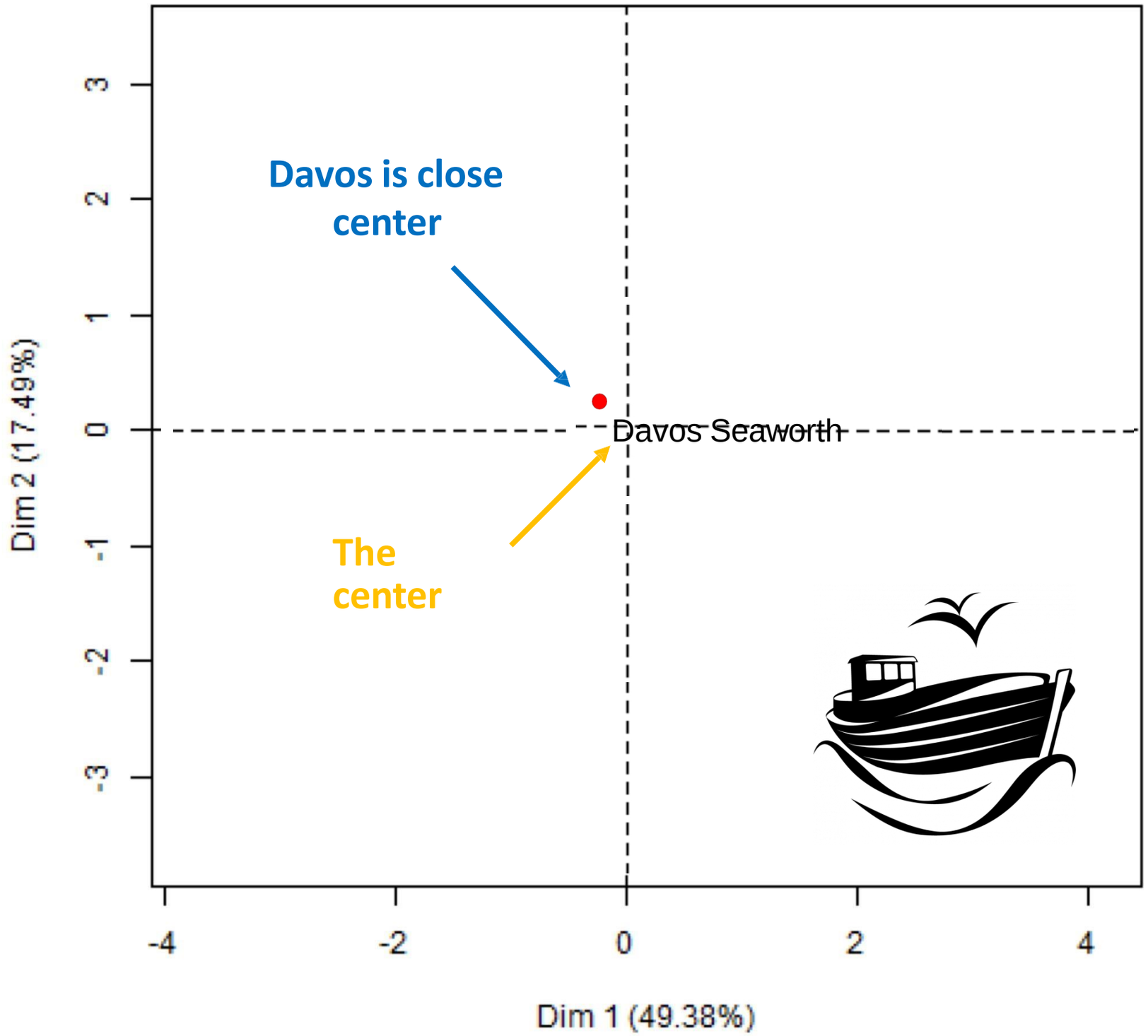
Dim 1

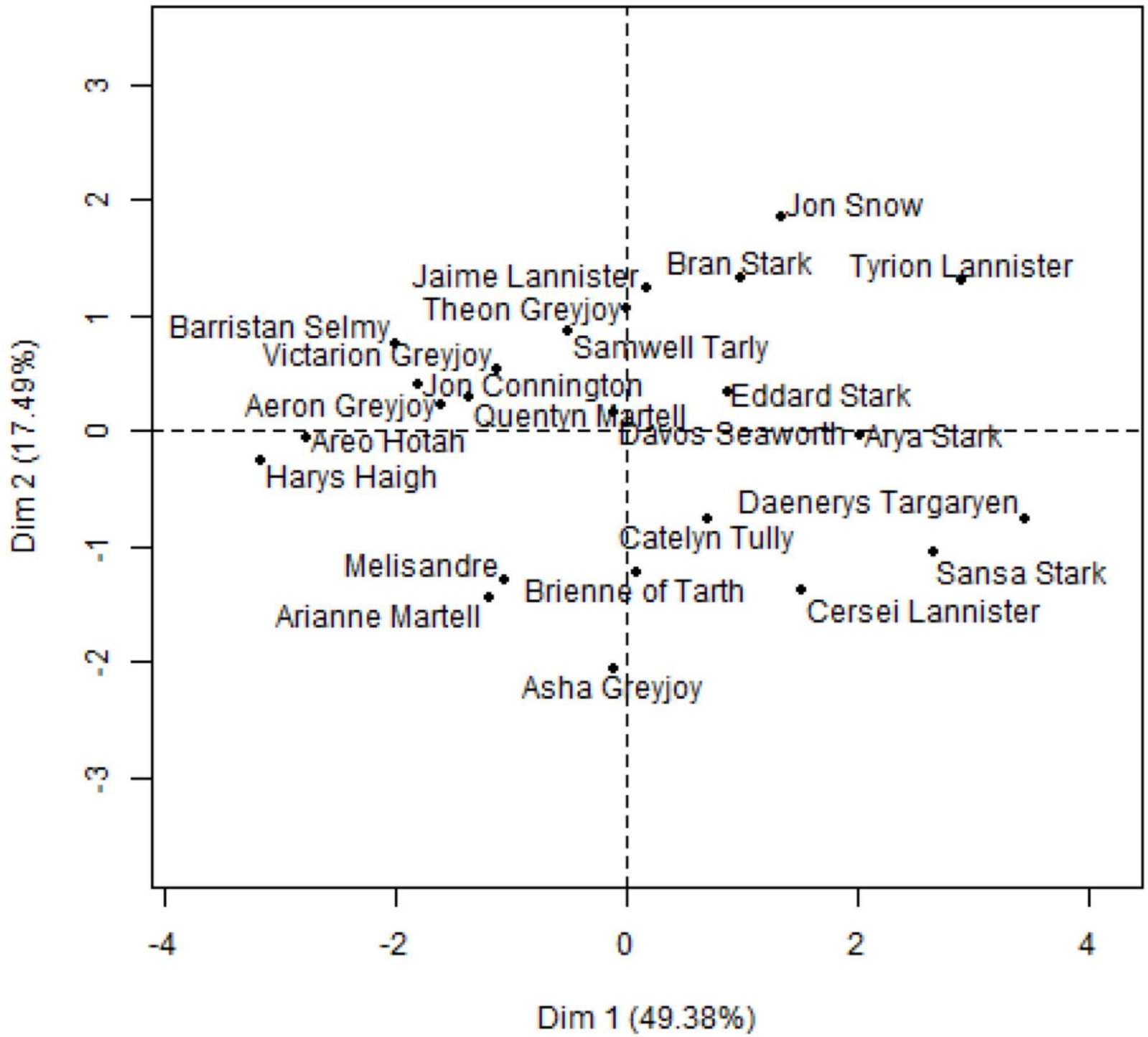

Poorly
represented
individual
=
"fuzzy
cos2
0
1

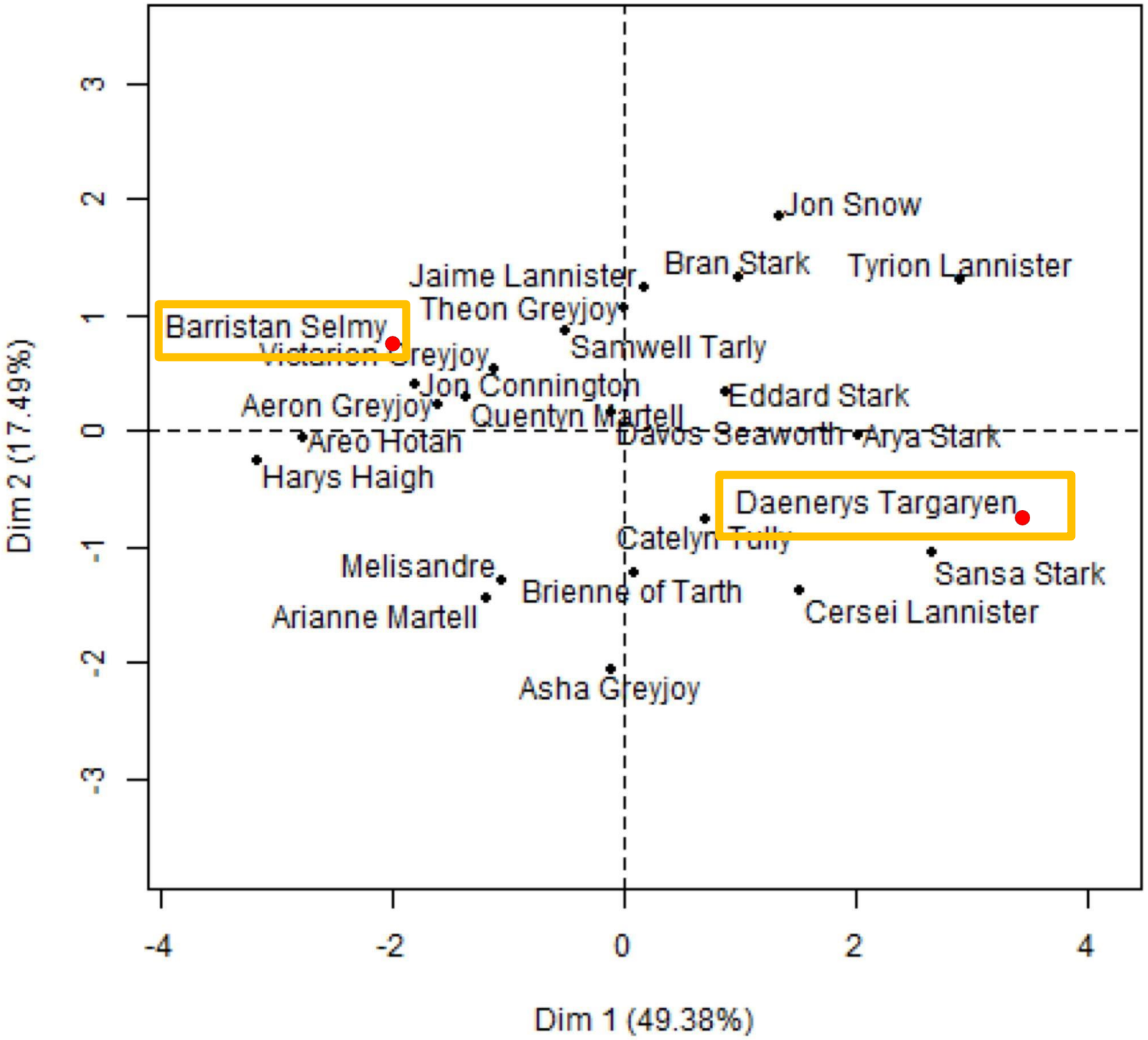

Dim 2 (17.49%)
3
2
1
0
-1
-2
-3
-4
Daenerys Targaryen
Sansa Stark
cos2
Daenerys
0
1
Sansa
0
1
Dim 1 (49.38%)

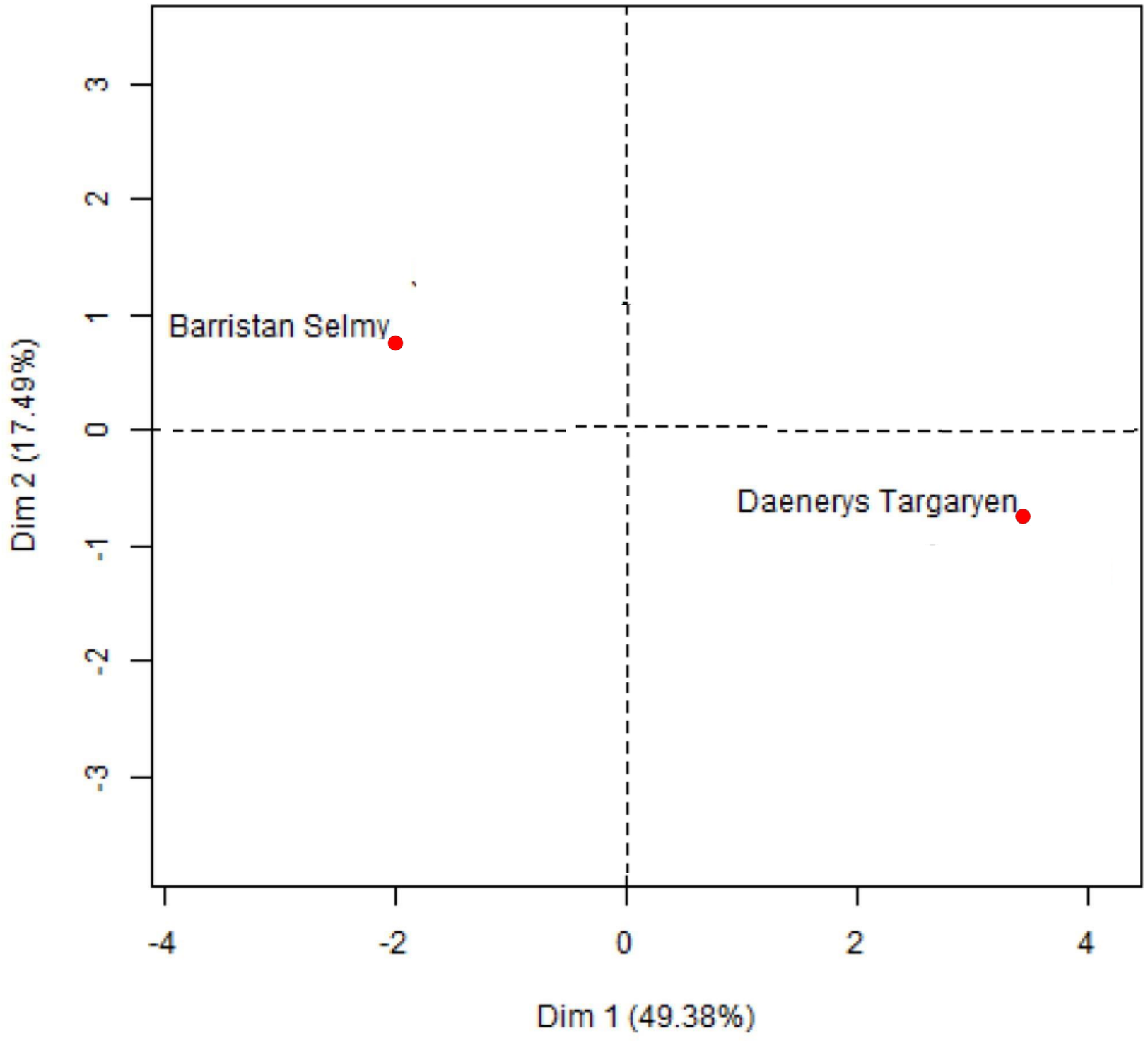

Jon Snow
Bran Stark
Tyrion Lannister
Jaime Lannister
Theon Greyjoy
Barristan Selmy
Samwell Tarly
Victarion Greyjoy
Aeron Greyjoy
Quentyn Martell
Davos Seaworth
Arya Stark
Areo Hotah
Harys Haigh
Daenerys Targaryen
Catelyn Tully
Sansa Stark
Melisandre
Brienne of Tarth
Cersei Lannister
Arianne Martell
Asha Greyjoy
Dim 2 (17.49%)
Dim 1 (49.38%)
3
2
1
0
-1
-2
-3
-4
-2
0
2
4

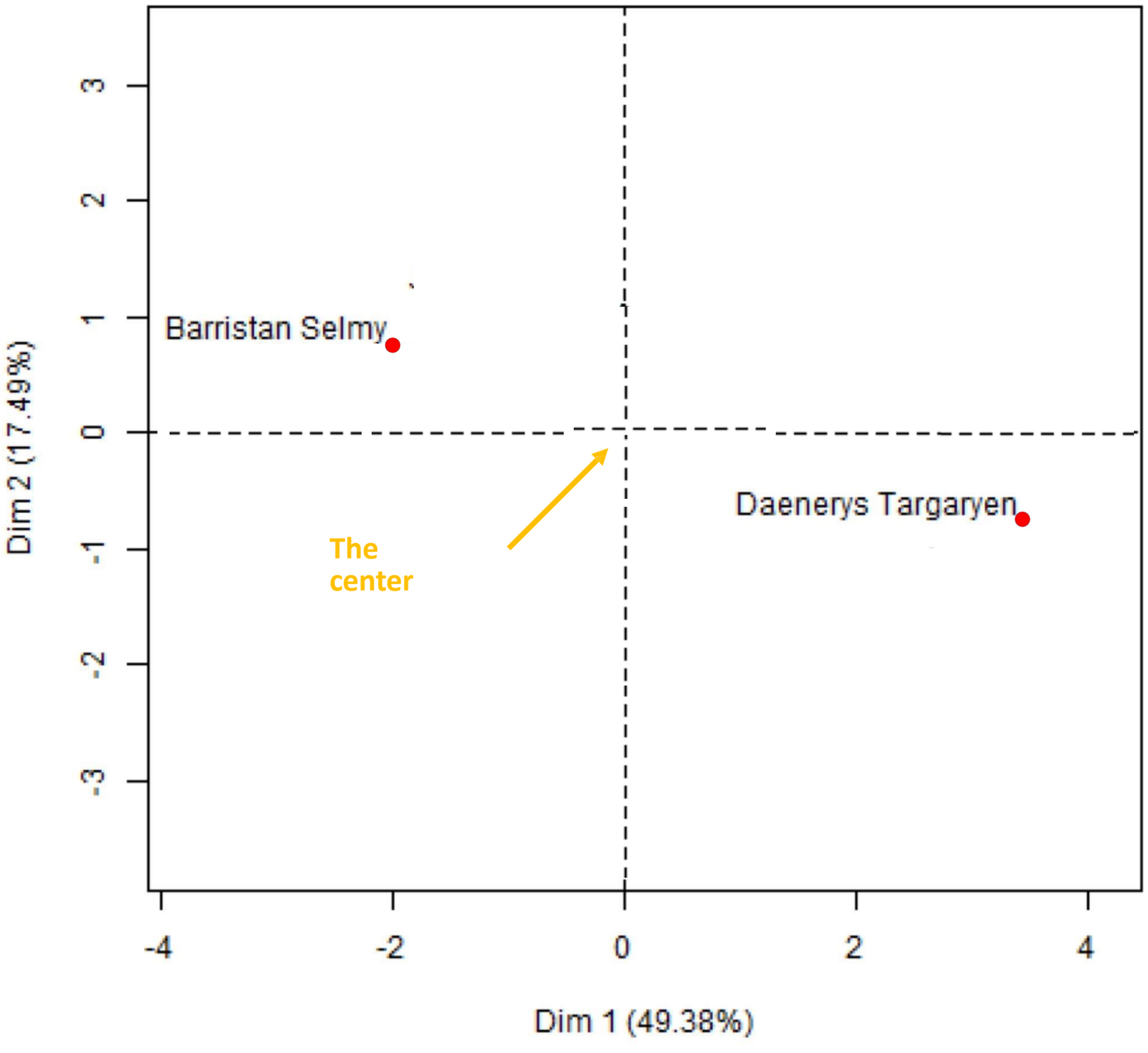

Davos Seaworth
Dim 2 (17.49%)
3
2
1
0
-1
-2
-3
-4
-2
0
2
4
Dim 1 (49.38%)

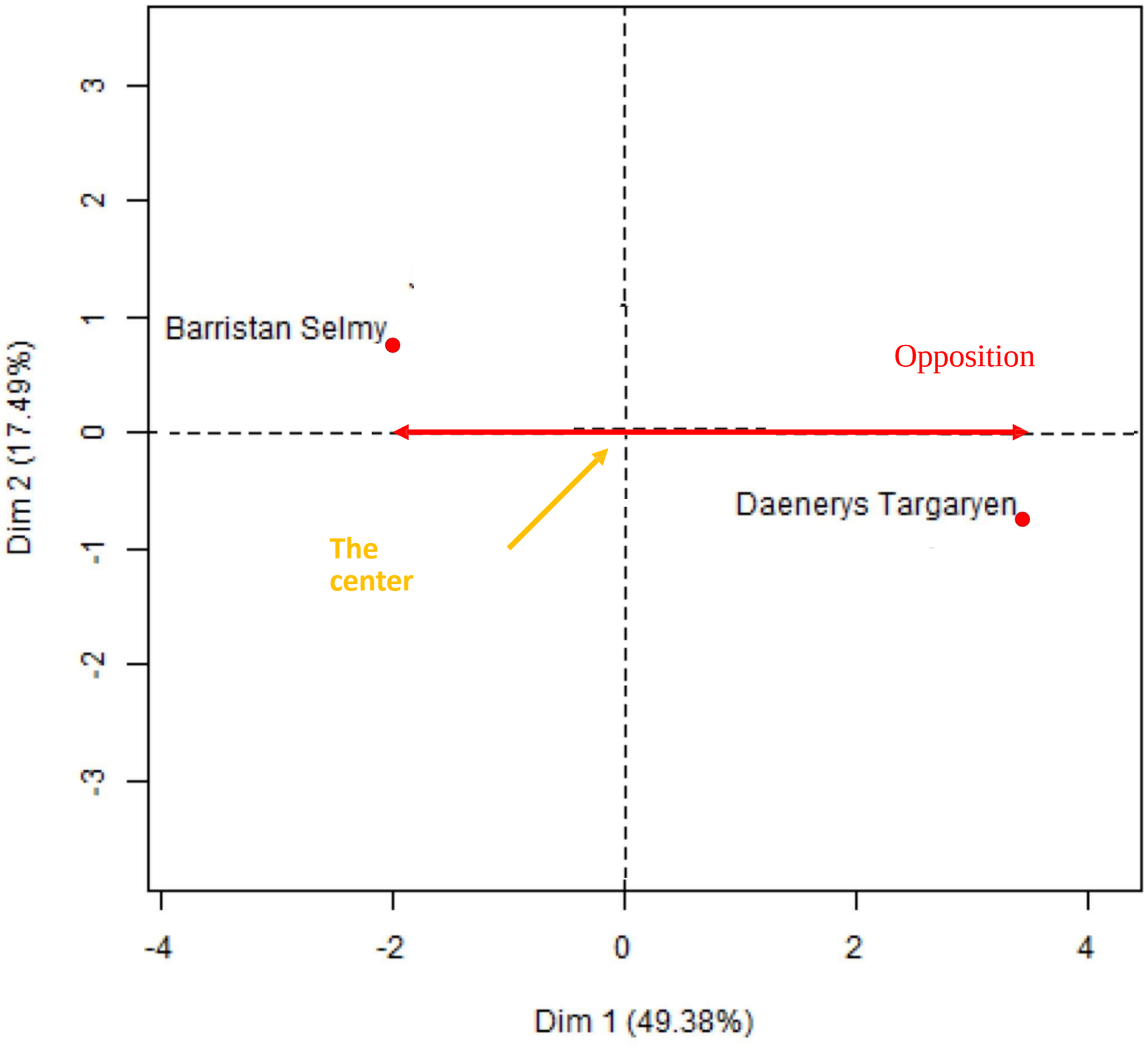

Davos is close center
Davos Seaworth
The center
Dim 2 (17.49%)
Dim 1 (49.38%)
3
2
1
0
-1
-2
-3
-4
-2
0
2
4

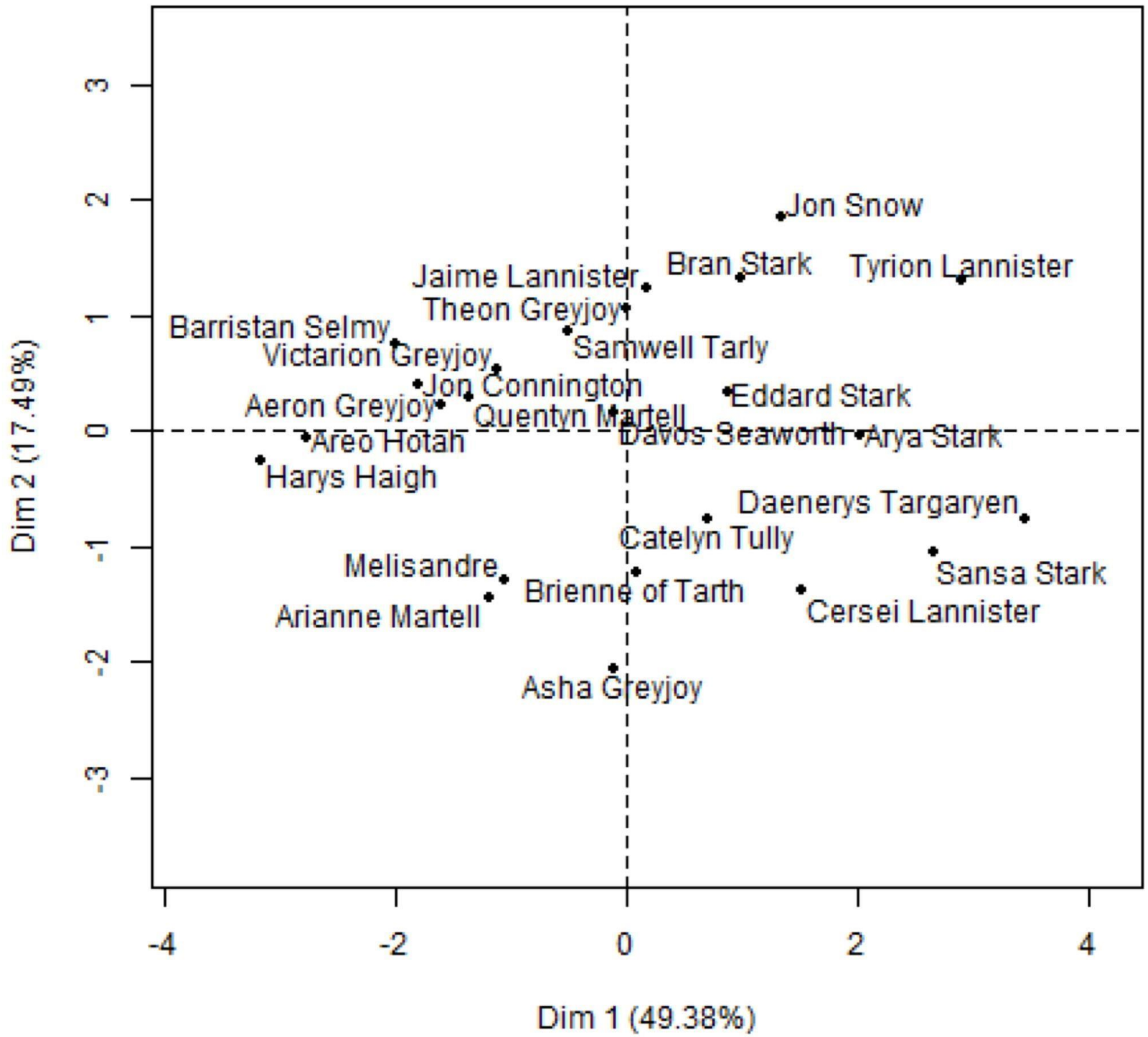

Jon Snow
Bran Stark
Tyrion Lannister
Jaime Lannister
Theon Greyjoy
Barristan Selmy
Samwell Tarly
Victarion Greyjoy
Jon Connington
Eddard Stark
Aeron Greyjoy
Quentyn Martell
Davos Seaworth
Arya Stark
Areo Hotah
Harys Haigh
Daenerys Targaryen
Catelyn Tully
Sansa Stark
Melisandre
Brienne of Tarth
Cersei Lannister
Arianne Martell
Asha Greyjoy
Dim 2 (17.49%)
Dim 1 (49.38%)
3
2
1
0
-1
-2
-3
-4
-2
0
2
4

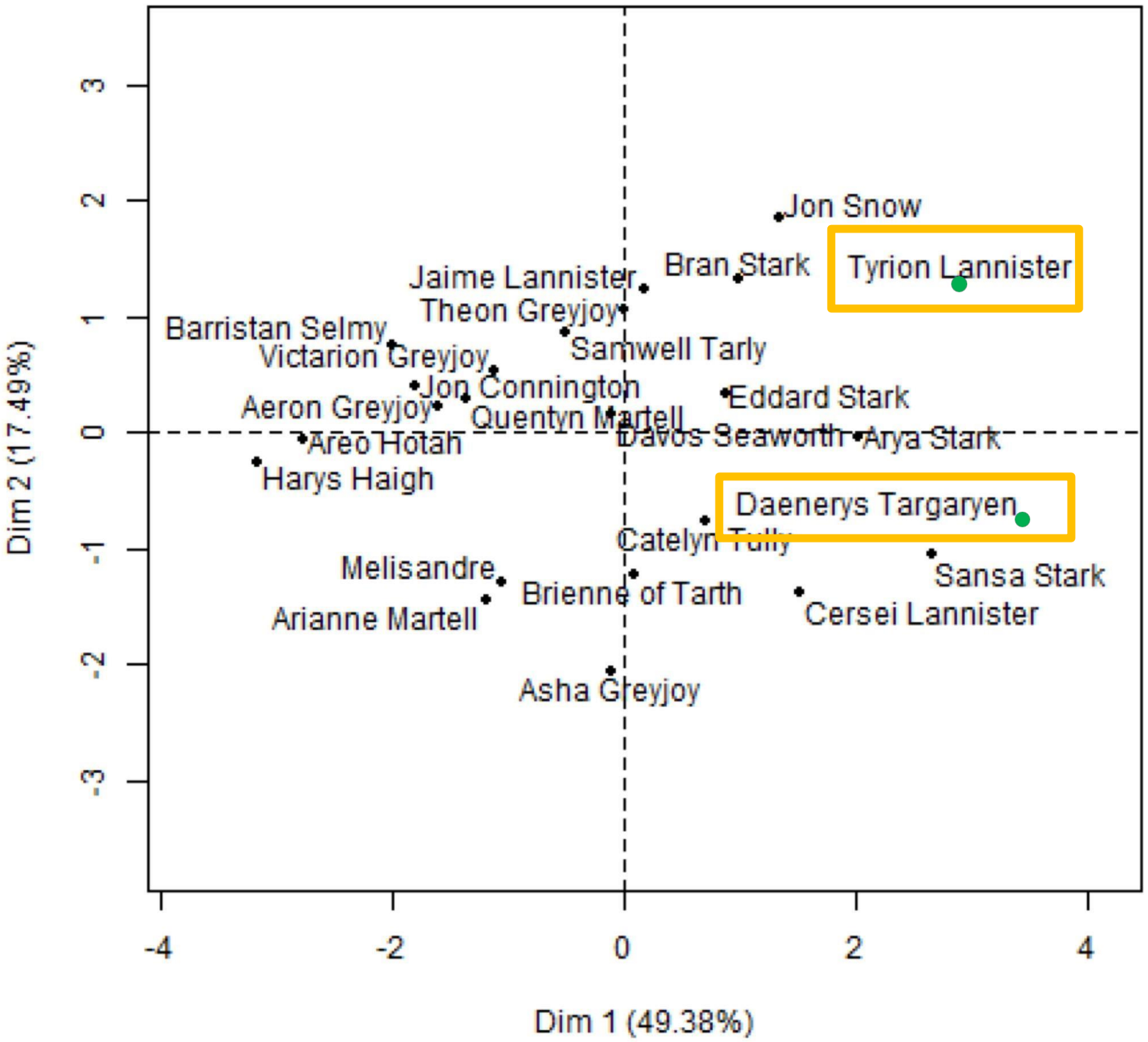

Jon Snow
Bran Stark
Tyrion Lannister
Jaime Lannister
Theon Greyjoy
Barristan Selmy
Samwell Tarly
Jon Connington
Eddard Stark
Aeron Greyjoy
Quentyn Martell
Davos Seaworth
Arya Stark
Areo Hotah
Harys Haigh
Daenerys Targaryen
Catelyn Tully
Sansa Stark
Melisandre
Brienne of Tarth
Arianne Martell
Cersei Lannister
Asha Greyjoy
3
2
1
0
-1
-2
-3
-4
-2
0
2
4
Dim 2 (17.49%)
Dim 1 (49.38%)

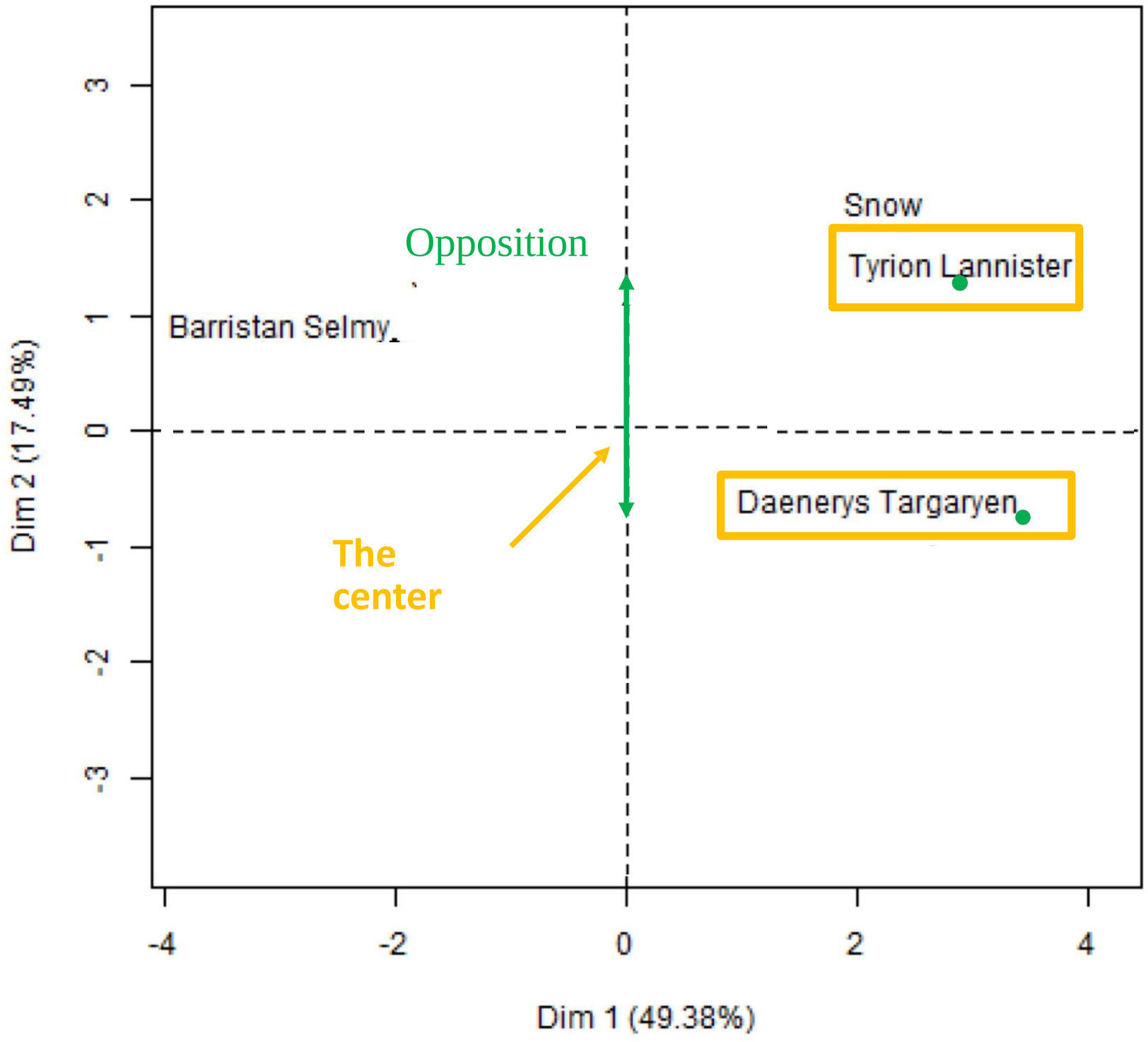

Barristan Selmy
Daenerys Targaryen
3
2
1
0
-1
-2
-3
-4
-2
0
2
4
Dim 2 (17.49%)
Dim 1 (49.38%)

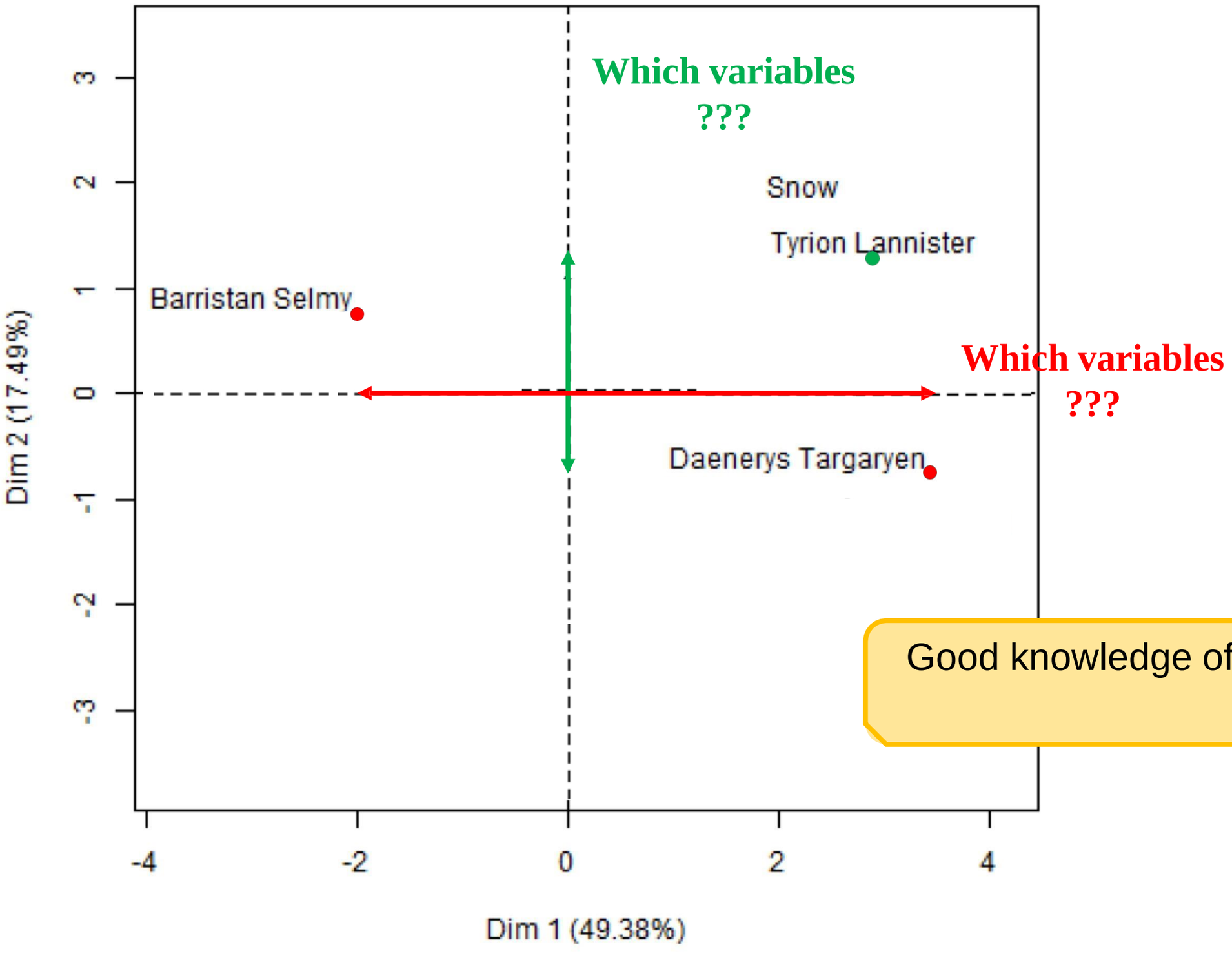

Barristan Selmy
Daenerys Targaryen
The center
Dim 2 (17.49%)
Dim 1 (49.38%)
3
2
1
0
-1
-2
-3
-4
-2
0
2
4


Barristan Selmy
Opposition
Daenerys Targaryen
The
center
Dim 2 (17.49%)
3
2
1
0
-1
-2
-3
-4
-2
0
2
4
Dim 1 (49.38%)

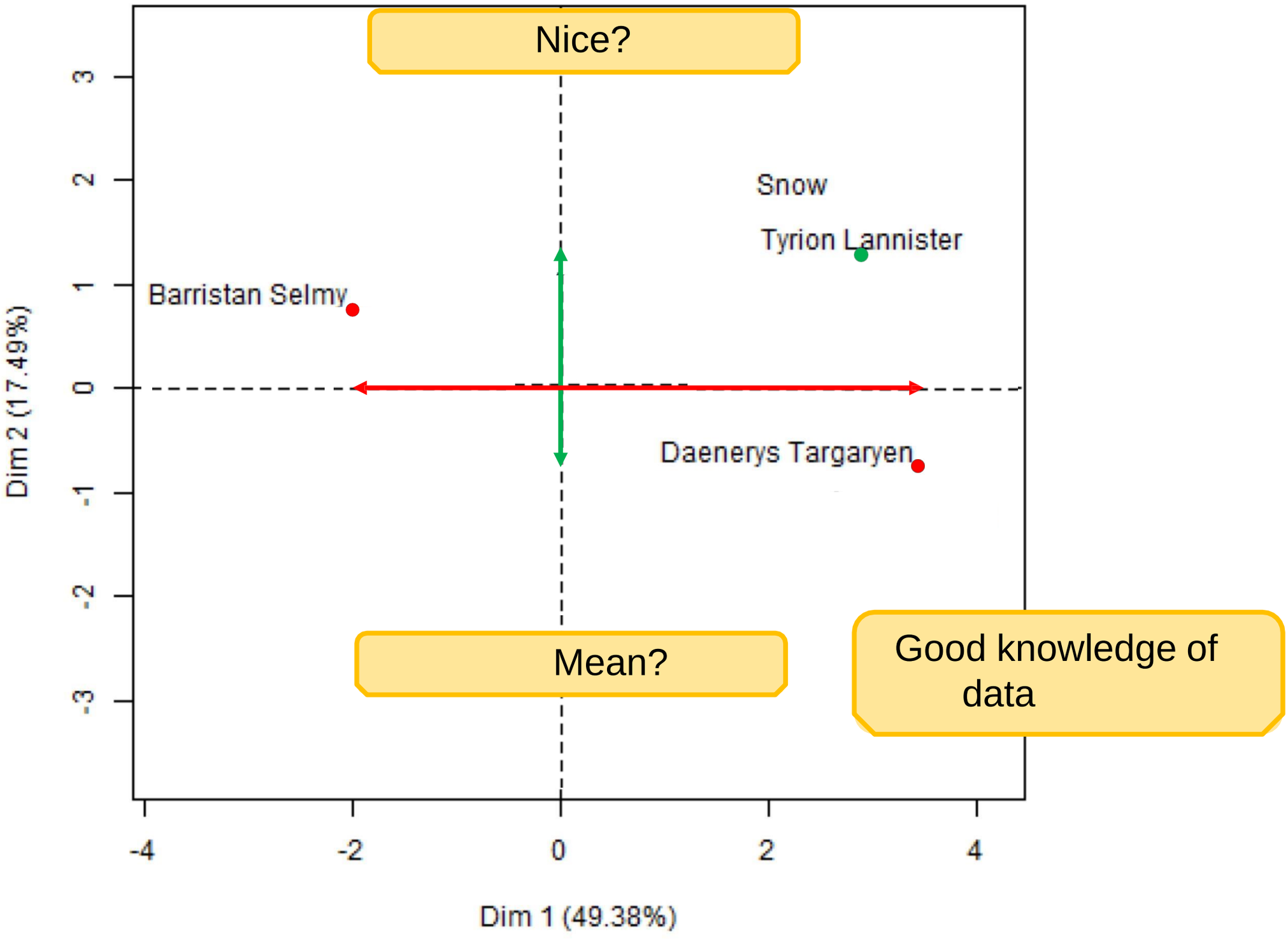

Jon Snow
Bran Stark
Tyrion Lannister
Jaime Lannister
Theon Greyjoy
Barristan Selmy
Samwell Tarly
Victarion Greyjoy
Jon Connington
Eddard Stark
Aeron Greyjoy
Quentyn Martell
Davos Seaworth
Arya Stark
Areo Hotah
Harys Haigh
Daenerys Targaryen
Catelyn Tully
Sansa Stark
Melisandre
Brienne of Tarth
Cersei Lannister
Arianne Martell
Asha Greyjoy
Dim 2 (17.49%)
Dim 1 (49.38%)
3
2
1
0
-1
-2
-3
-4
-2
0
2
4

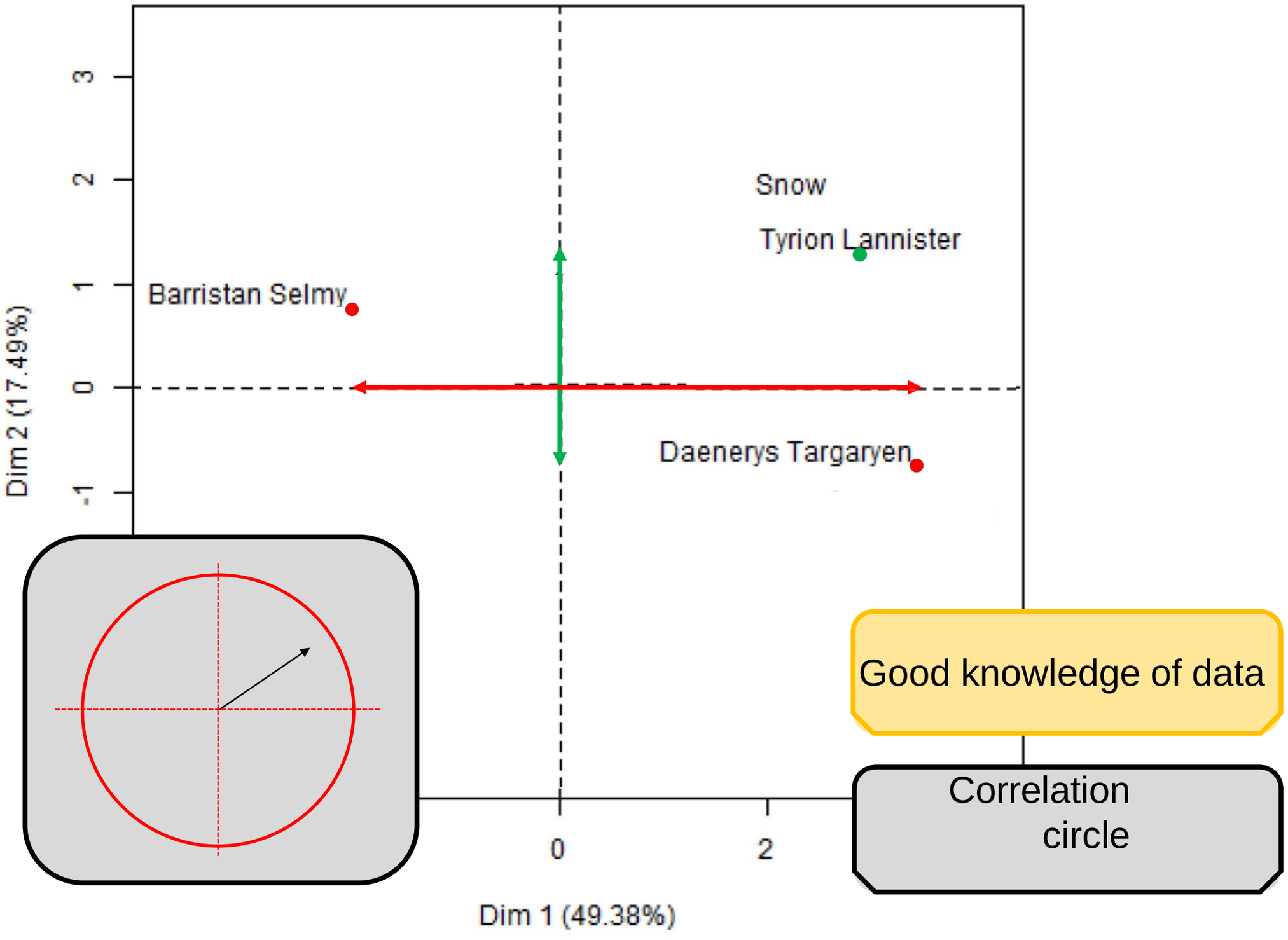

Jon Snow
Bran Stark
Tyrion Lannister
Jaime Lannister
Theon Greyjoy
Barristan Selmy
Samwell Tarly
Victarion Greyjoy
Jon Connington
Eddard Stark
Aeron Greyjoy
Quentyn Martell
Davos Seaworth
Arya Stark
Areo Hotah
Harys Haigh
Daenerys Targaryen
Catelyn Tully
Sansa Stark
Melisandre
Brienne of Tarth
Arianne Martell
Cersei Lannister
Asha Greyjoy
3
2
1
0
-1
-2
-3
-4
-2
0
2
4
Dim 2 (17.49%)
Dim 1 (49.38%)


Opposition
Barristan Selmy
Snow
Tyrion Lannister
Daenerys Targaryen
The
center
Dim 2 (17.49%)
3
2
1
0
-1
-2
-3
-4
-2
0
2
4
Dim 1 (49.38%)



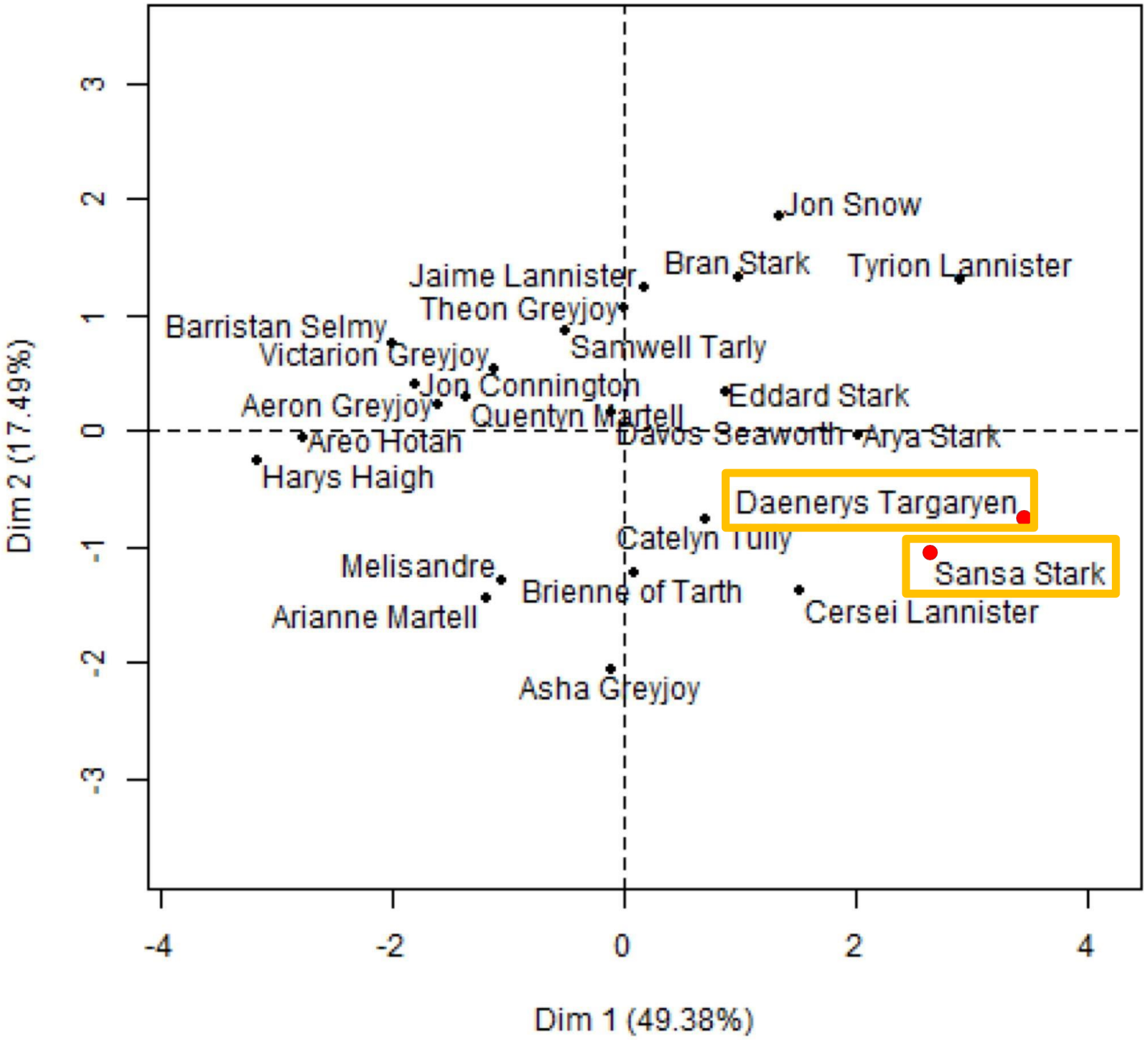
Good knowledge of data

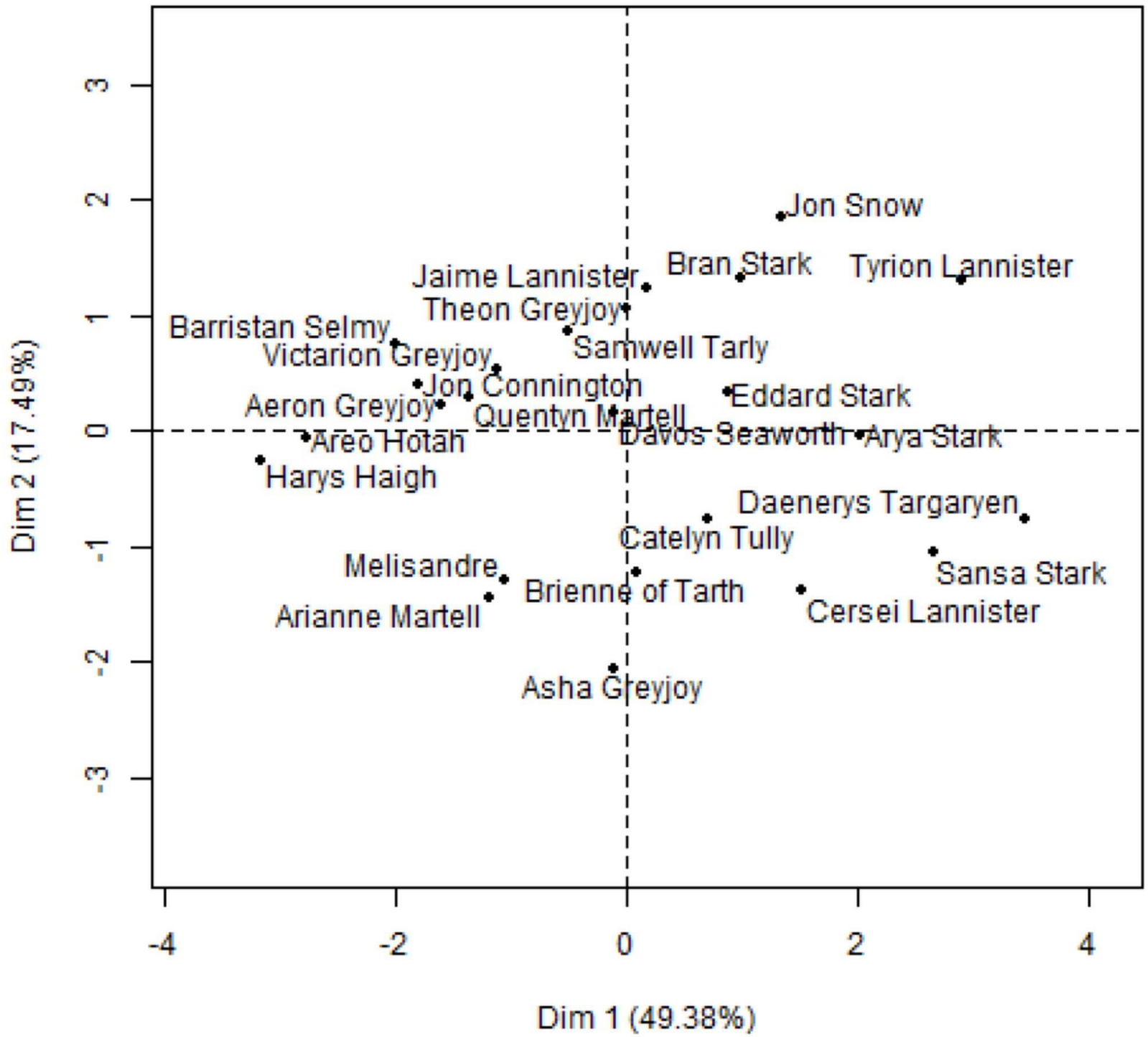

Nice?
Snow
Tyrion Lannister
Barristan Selmy
Daenerys Targaryen
Mean?
Good knowledge of data
Dim 2 (17.49%)
Dim 1 (49.38%)
3
2
1
0
-1
-2
-3
-4
-2
0
2
4



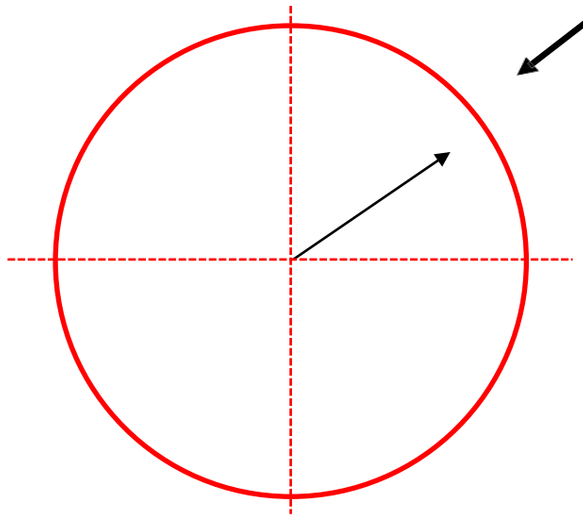
Good knowledge of data

Correlation circle

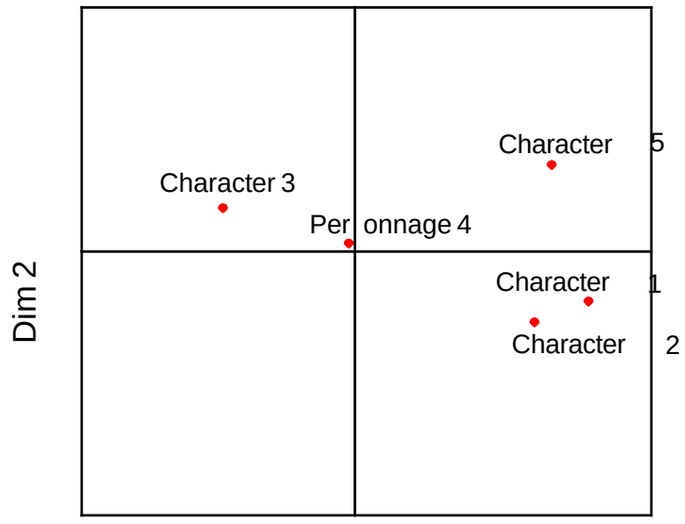

Jon Snow
Bran Stark
Tyrion Lannister
Jaime Lannister
Theon Greyjoy
Barristan Selmy
Samwell Tarly
Victarion Greyjoy
Jon Connington
Eddard Stark
Aeron Greyjoy
Quentyn Martell
Davos Seaworth
Arya Stark
Areo Hotah
Harys Haigh
Daenerys Targaryen
Catelyn Tully
Sansa Stark
Melisandre
Brienne of Tarth
Arianne Martell
Cersei Lannister
Asha Greyjoy
3
2
1
0
-1
-2
-3
-4
-2
0
2
4
Dim 2 (17.49%)
Dim 1 (49.38%)


Jon Snow
Bran Stark
Tyrion Lannister
Jaime Lannister
Theon Greyjoy
Barristan Selmy
Samwell Tarly
Victarion Greyjoy
Jon Connington
Eddard Stark
Aeron Greyjoy
Quentyn Martell
Davos Seaworth
Arya Stark
Areo Hotah
Harys Haigh
Daenerys Targaryen
Catelyn Tully
Sansa Stark
Melisandre
Brienne of Tarth
Cersei Lannister
Arianne Martell
Asha Greyjoy
Dim 2 (17.49%)
Dim 1 (49.38%)
3
2
1
0
-1
-2
-3
-4
-2
0
2
4

ACP



orient



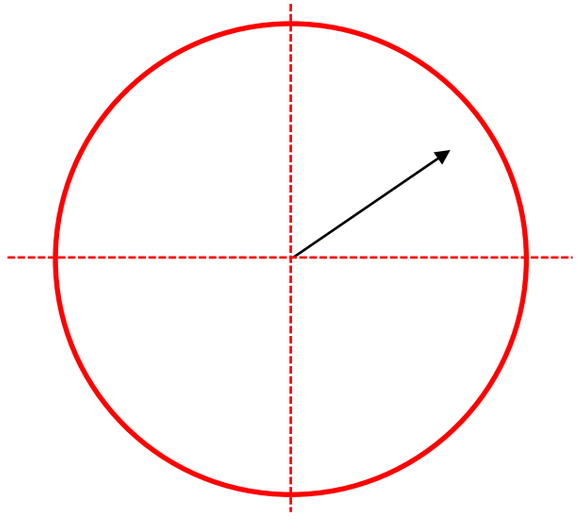
Correlation circle

Plot of individuals

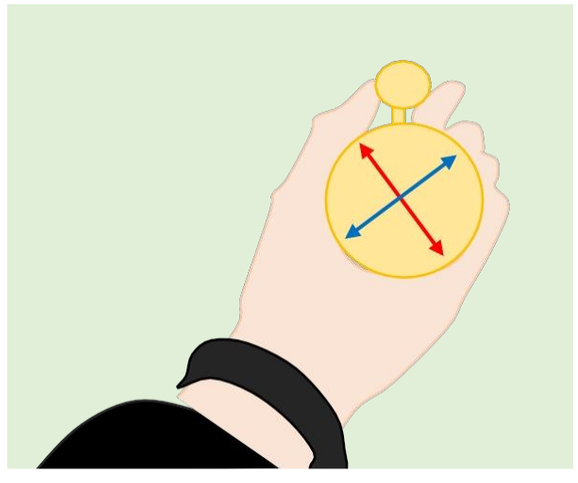
Compass

orient

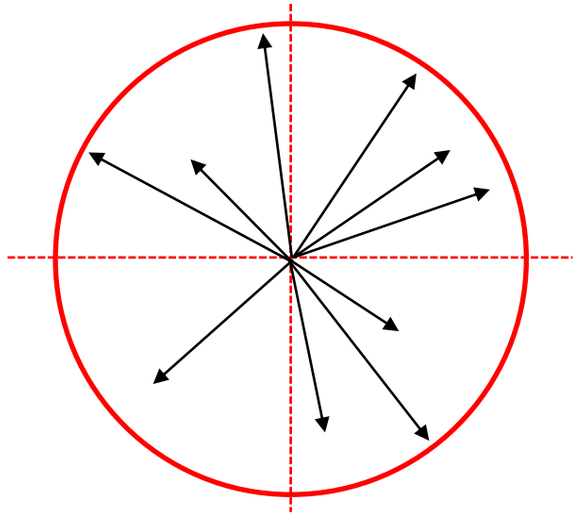
Photo



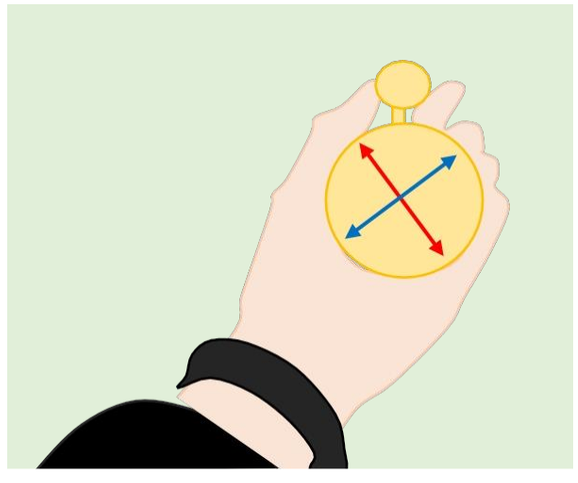
Correlation circle



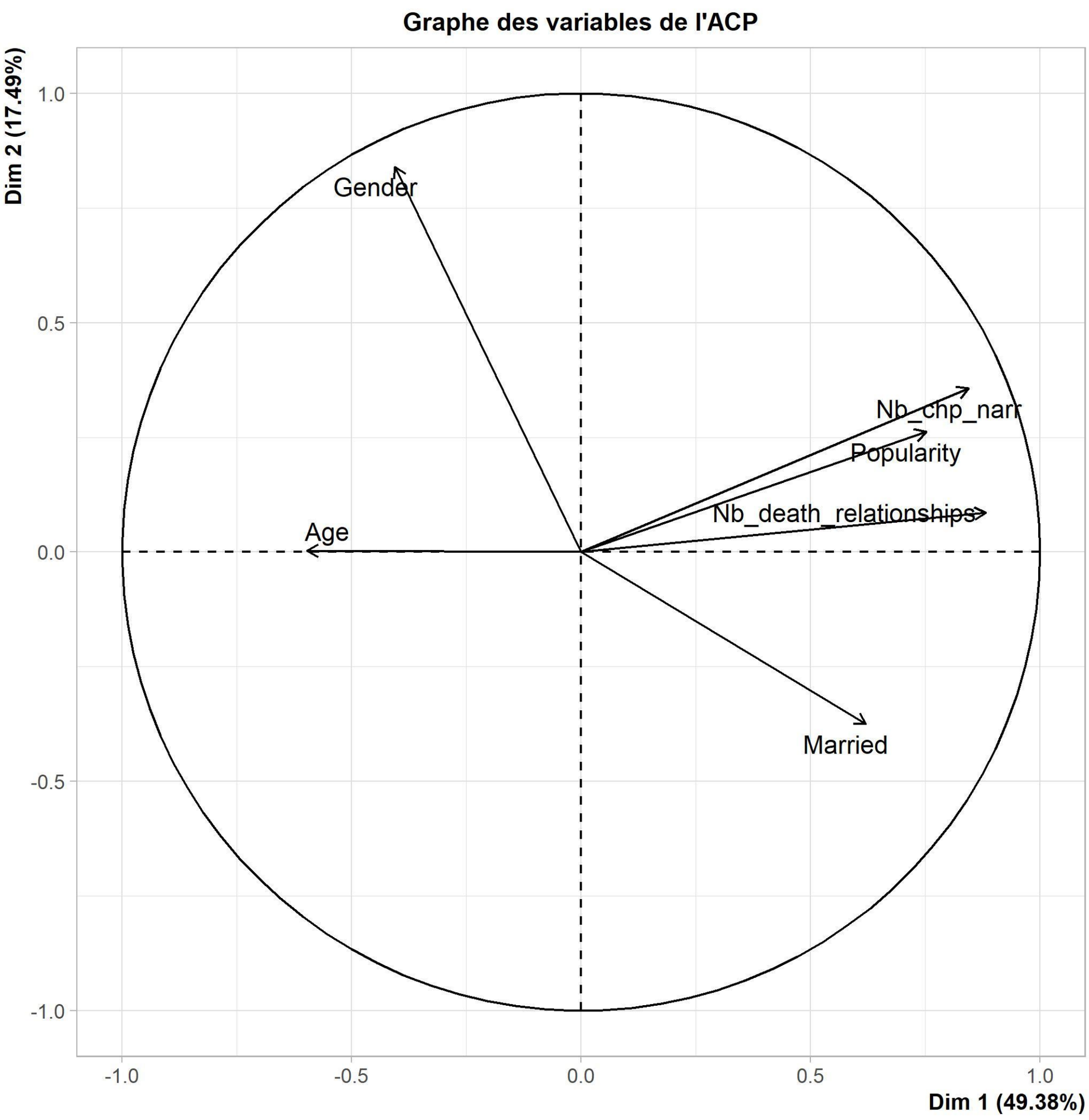
Compass

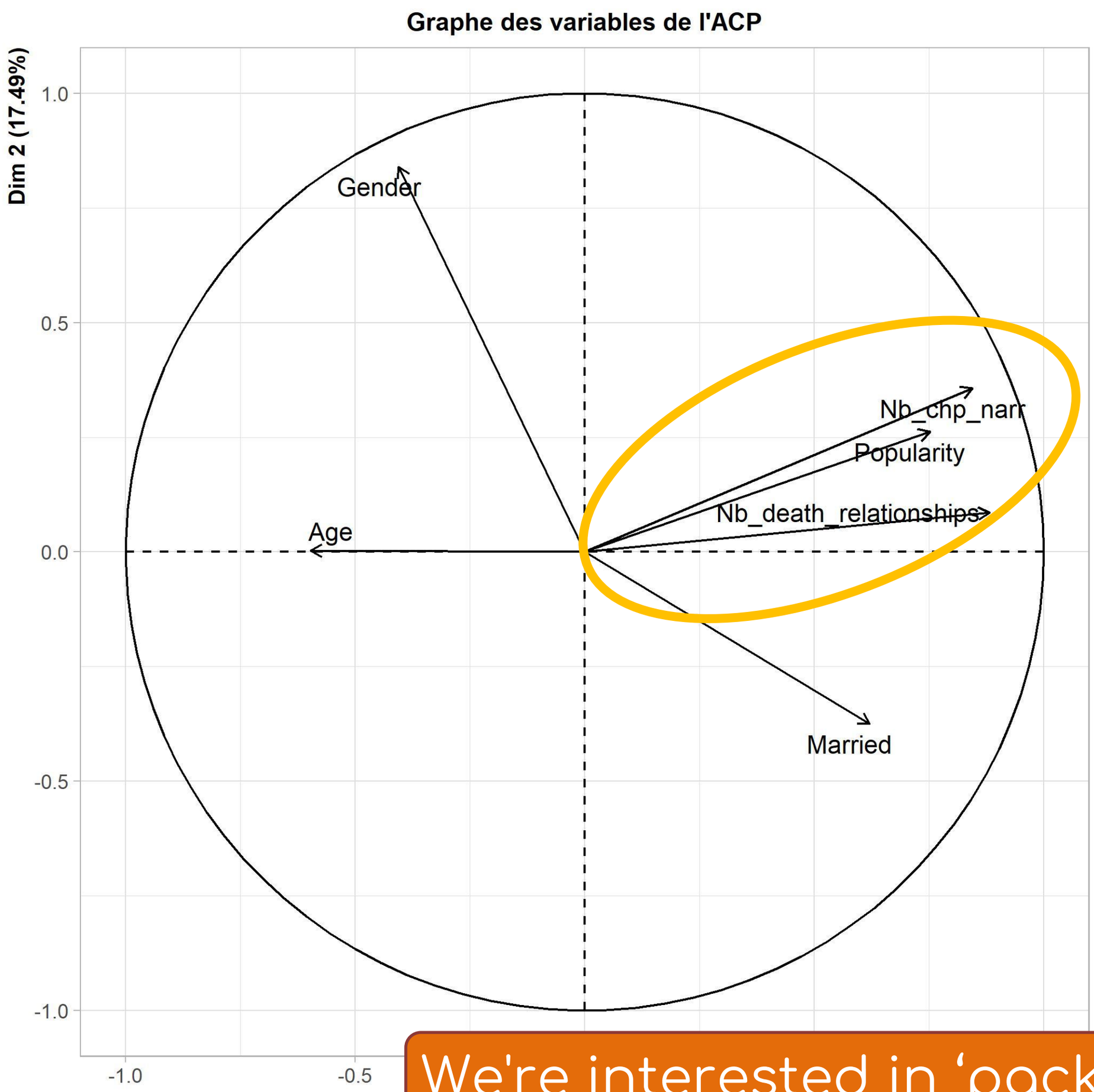


Correlation circle

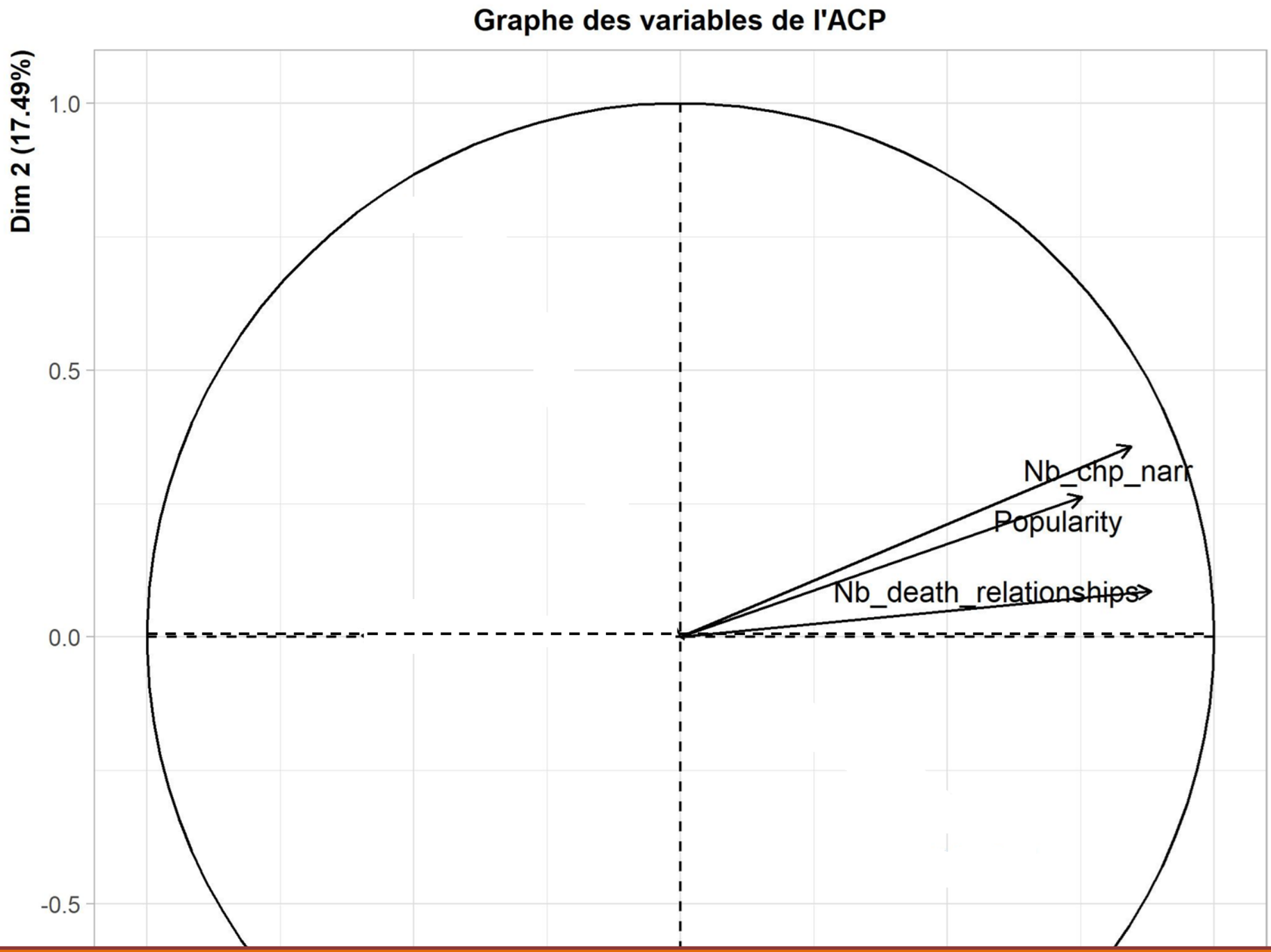


Compass


Graphe des variables de l'ACP
Dim 2 (17.49%)
1.0
0.5
0.0
-0.5
-1.0
Gender
Age
Nb_chp_narr
Popularity
Nb_death_relationships
Married
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (49.38%)



We're interested in ‘packages’.



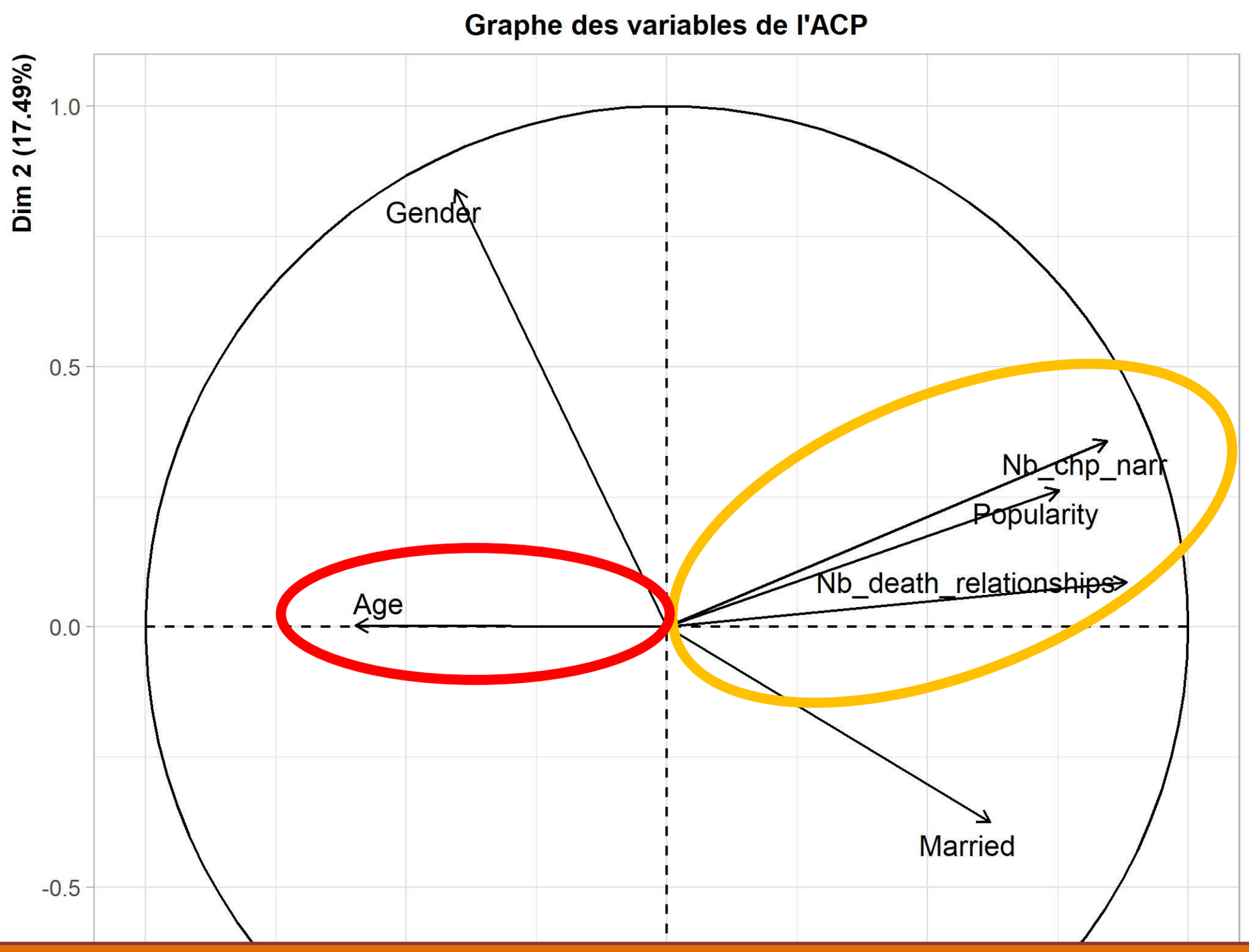
Number of chapters narrated

Popular

deaths in their relationships

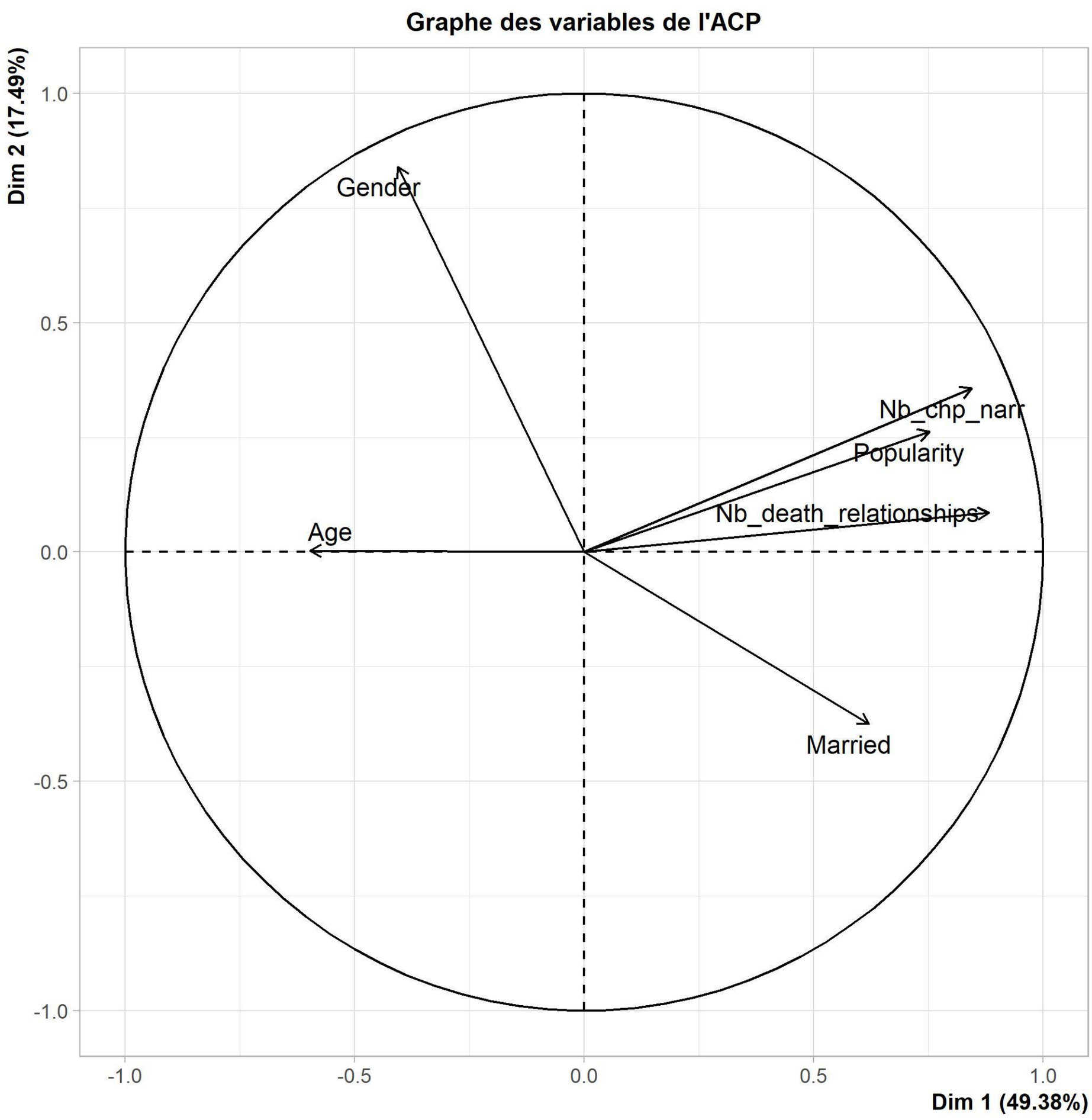
Parallel or almost-parallel arrows pointing in the same direction - these are related variables.

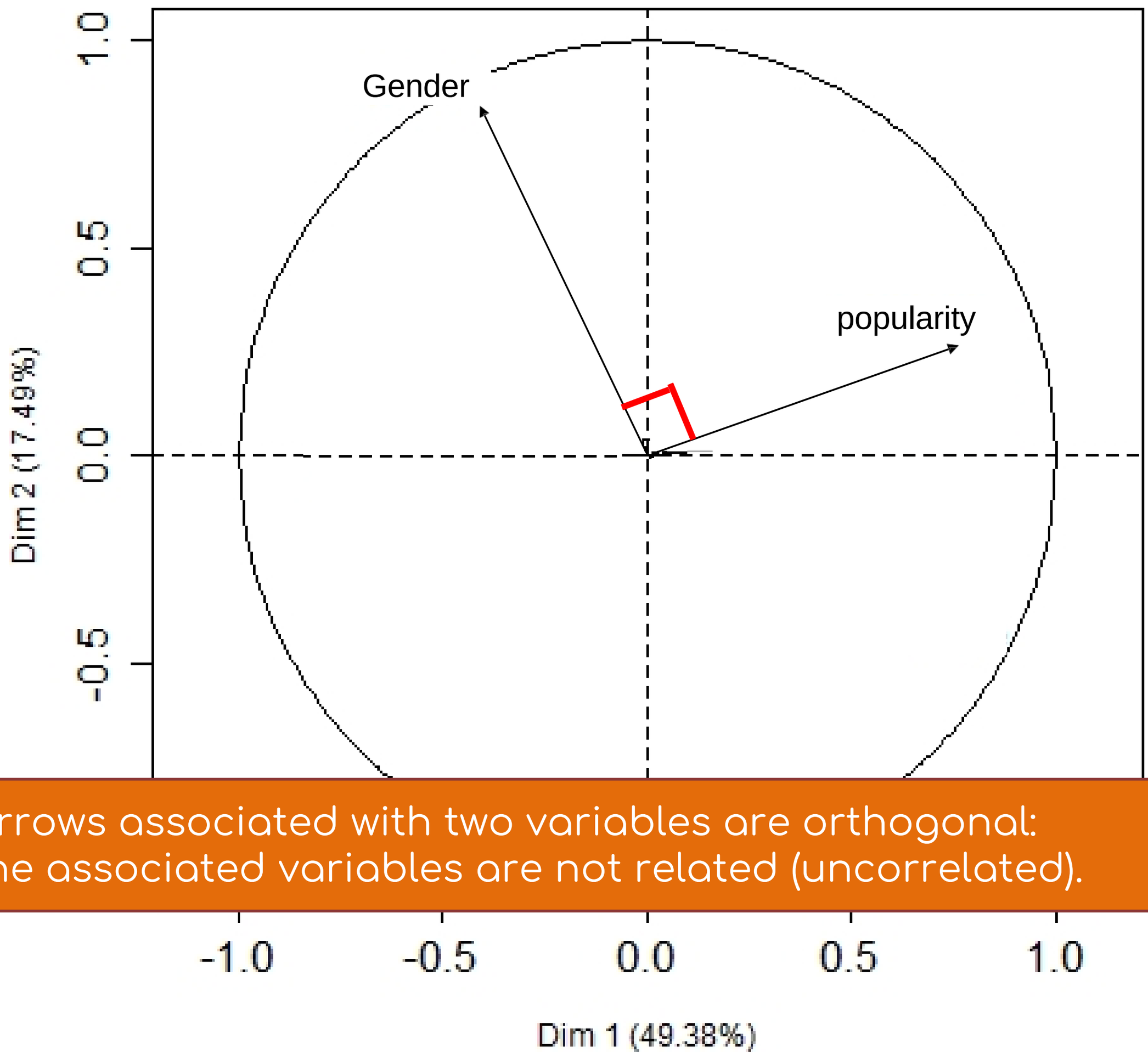
Positive correlation: when the values of one variable increase, those of the other decrease.



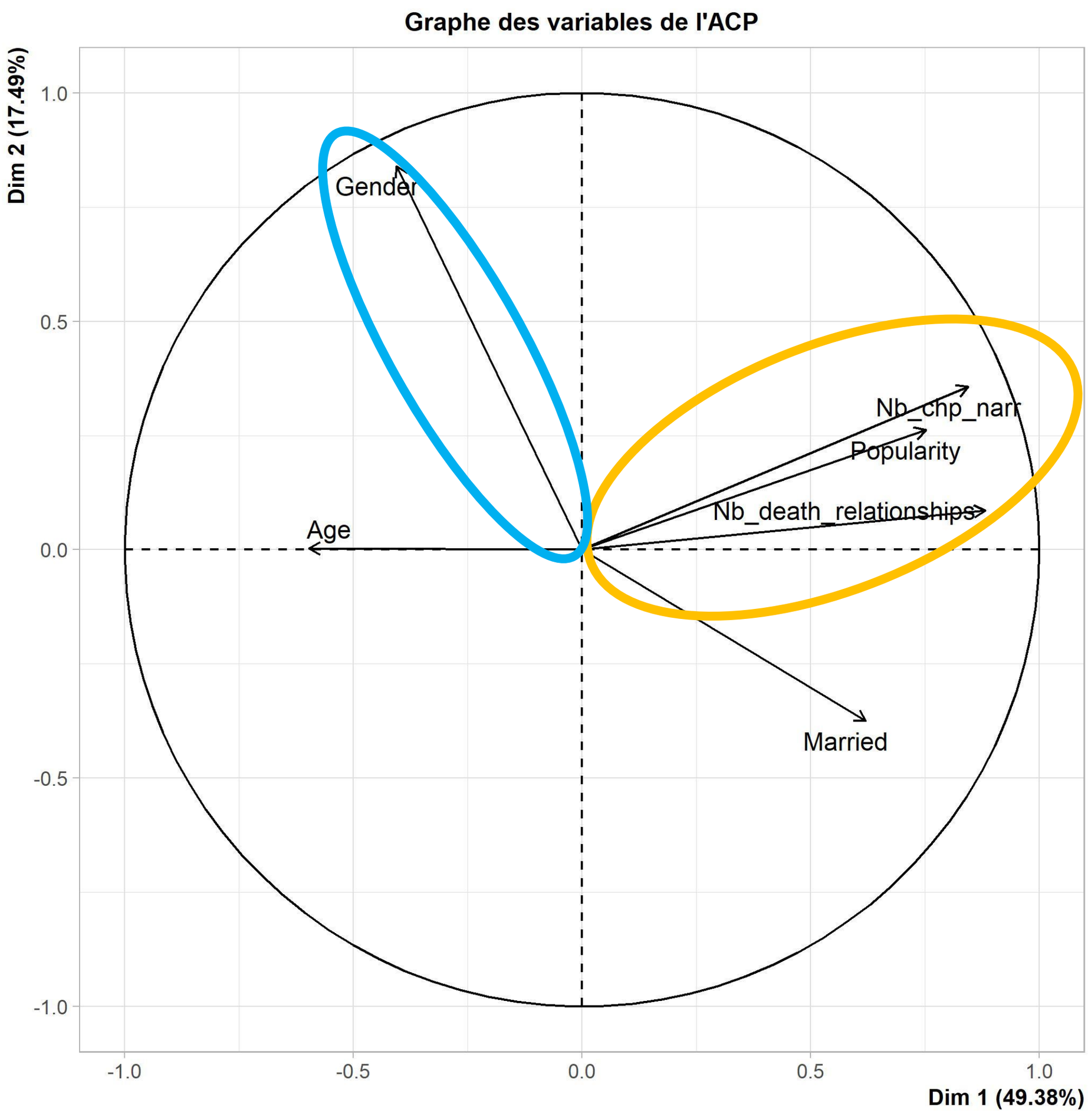
Parallel or almost parallel arrows pointing in opposite directions are also related variables.

Negative correlation: when the values of one variable increase, those of the other decrease.

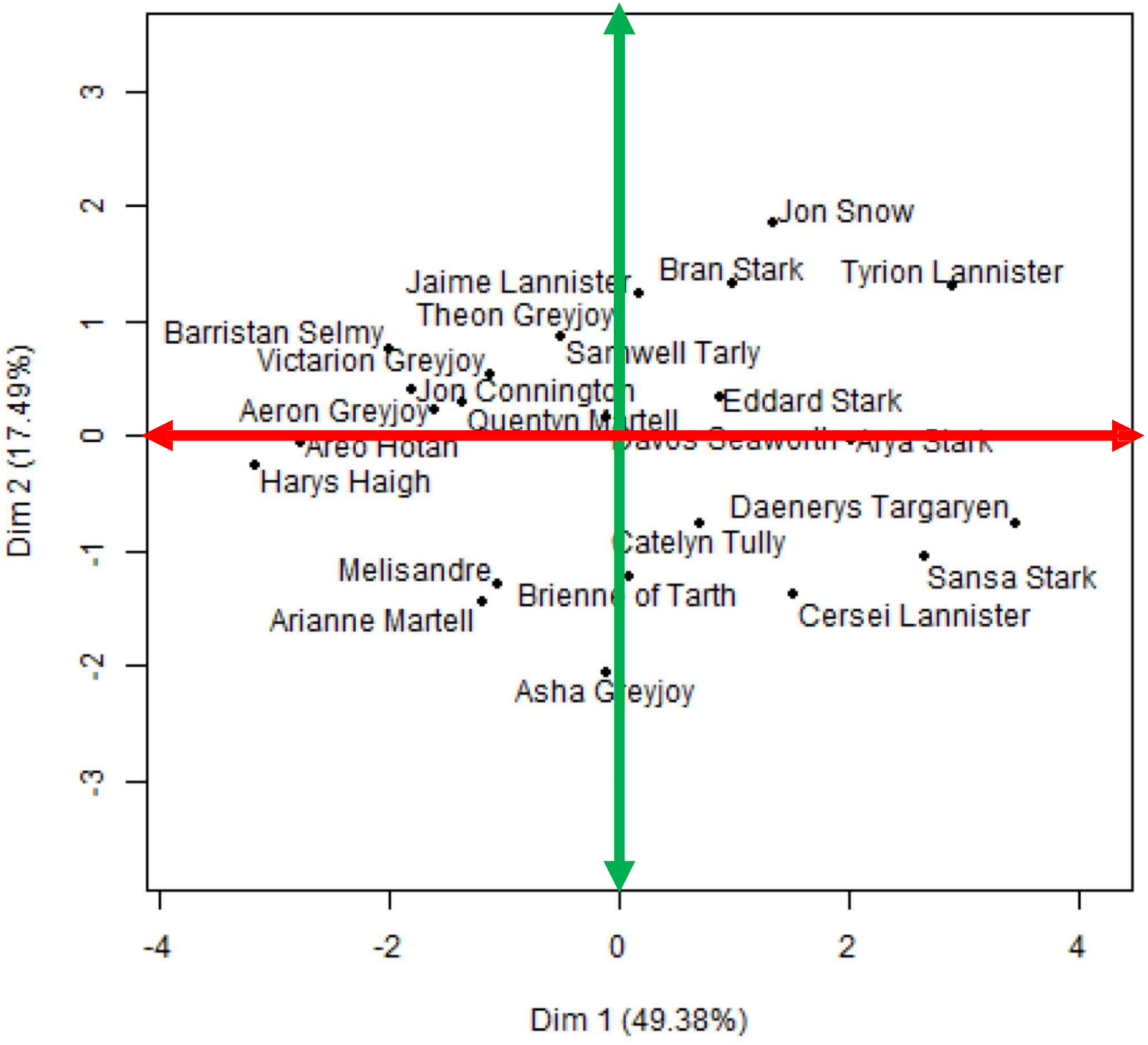

Graphe des variables de l'ACP
Dim 2 (17.49%)
1.0
0.5
0.0
-0.5
-1.0
Gender
Nb_chp_narr
Popularity
Nb_death_relationships
Age
Married
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (49.38%)



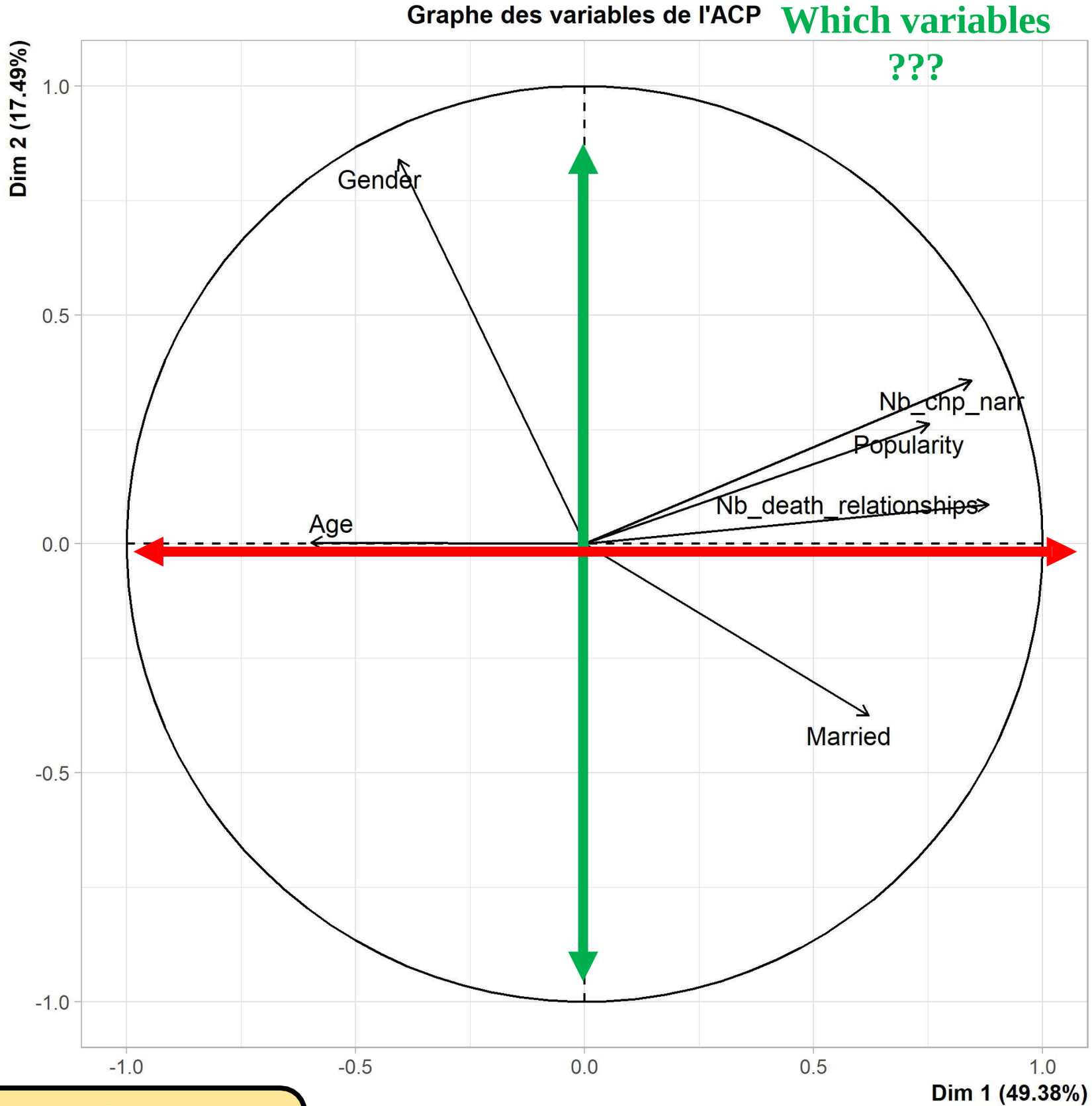
Arrows associated with two variables are orthogonal:
the associated variables are not related (uncorrelated).


Graphe des variables de l'ACP
Dim 2 (17.49%)
1.0
0.5
0.0
-0.5
-1.0
Gender
Age
Nb_chp_narr
Popularity
Nb_death_relationships
Married
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (49.38%)

Which variables ???



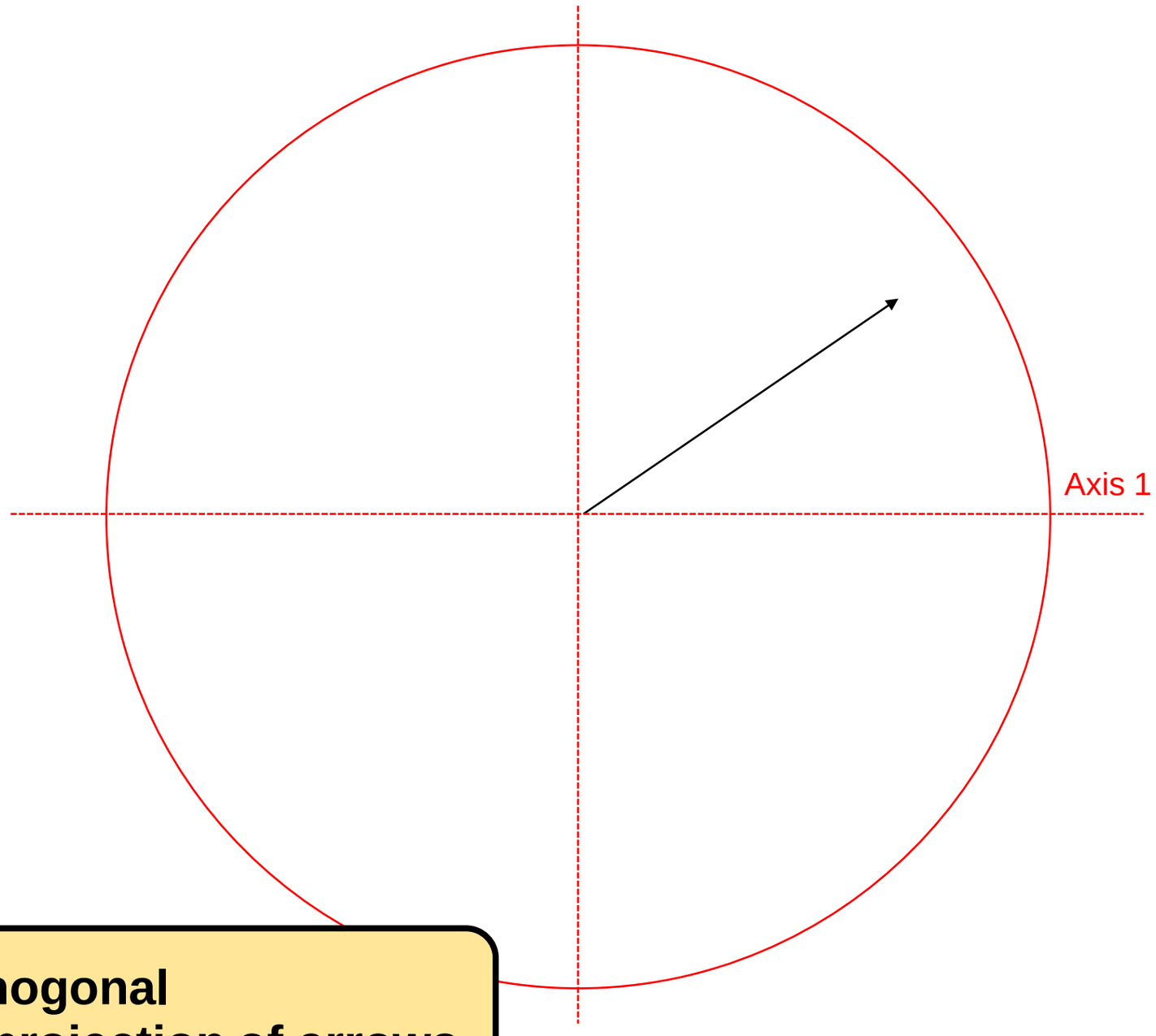
Which variables ???



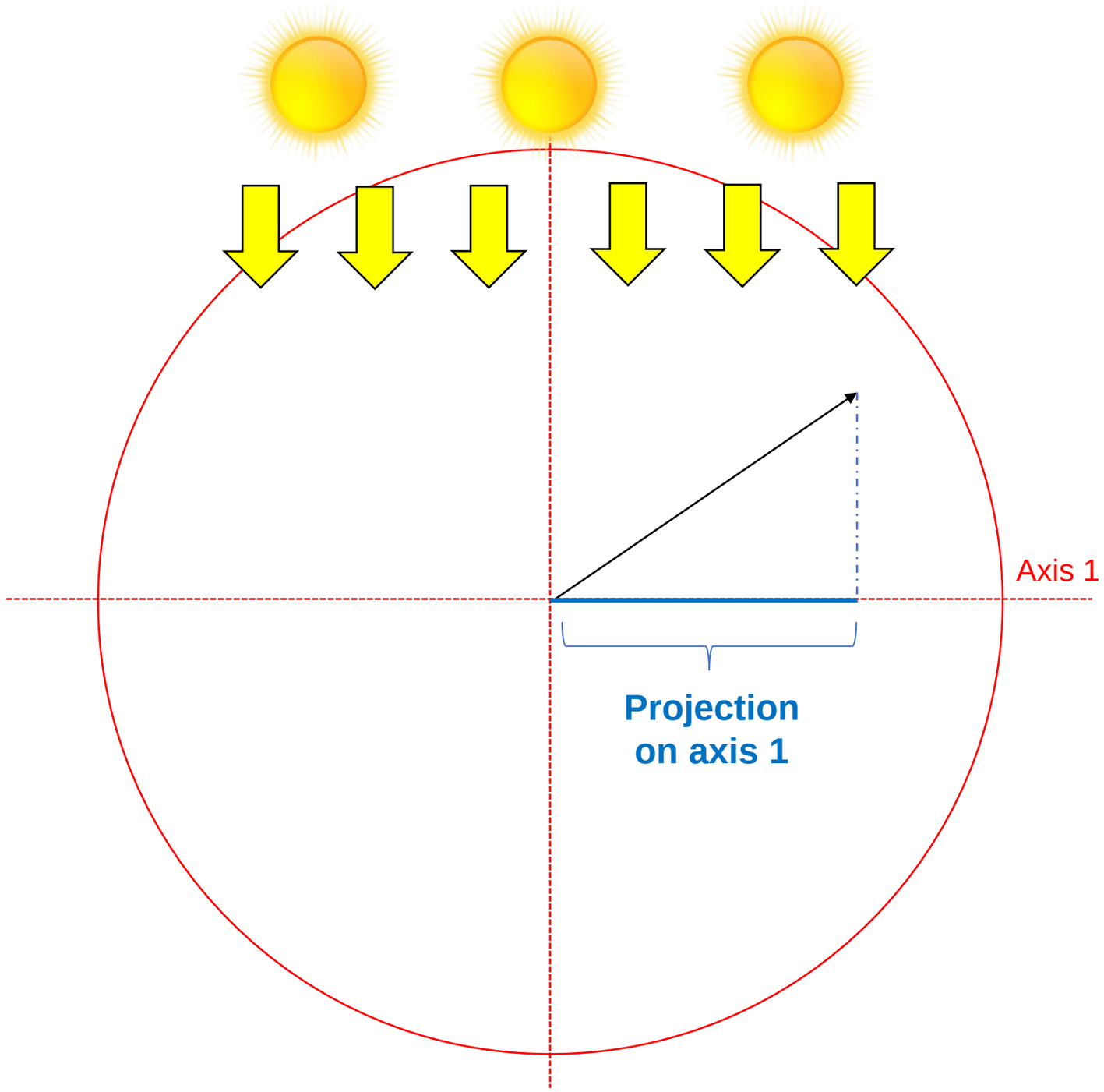
Which variables ???

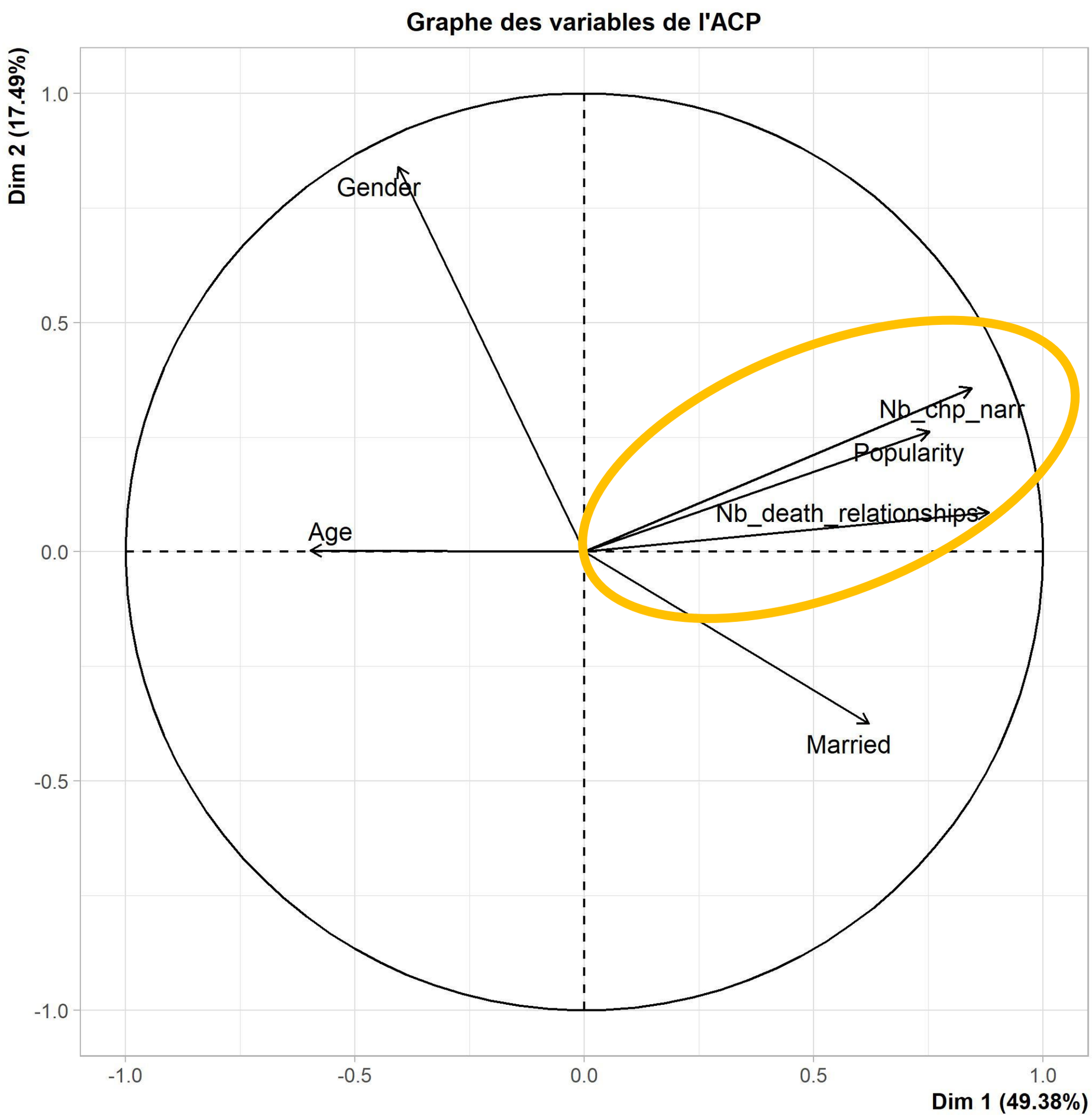
Which variables ???

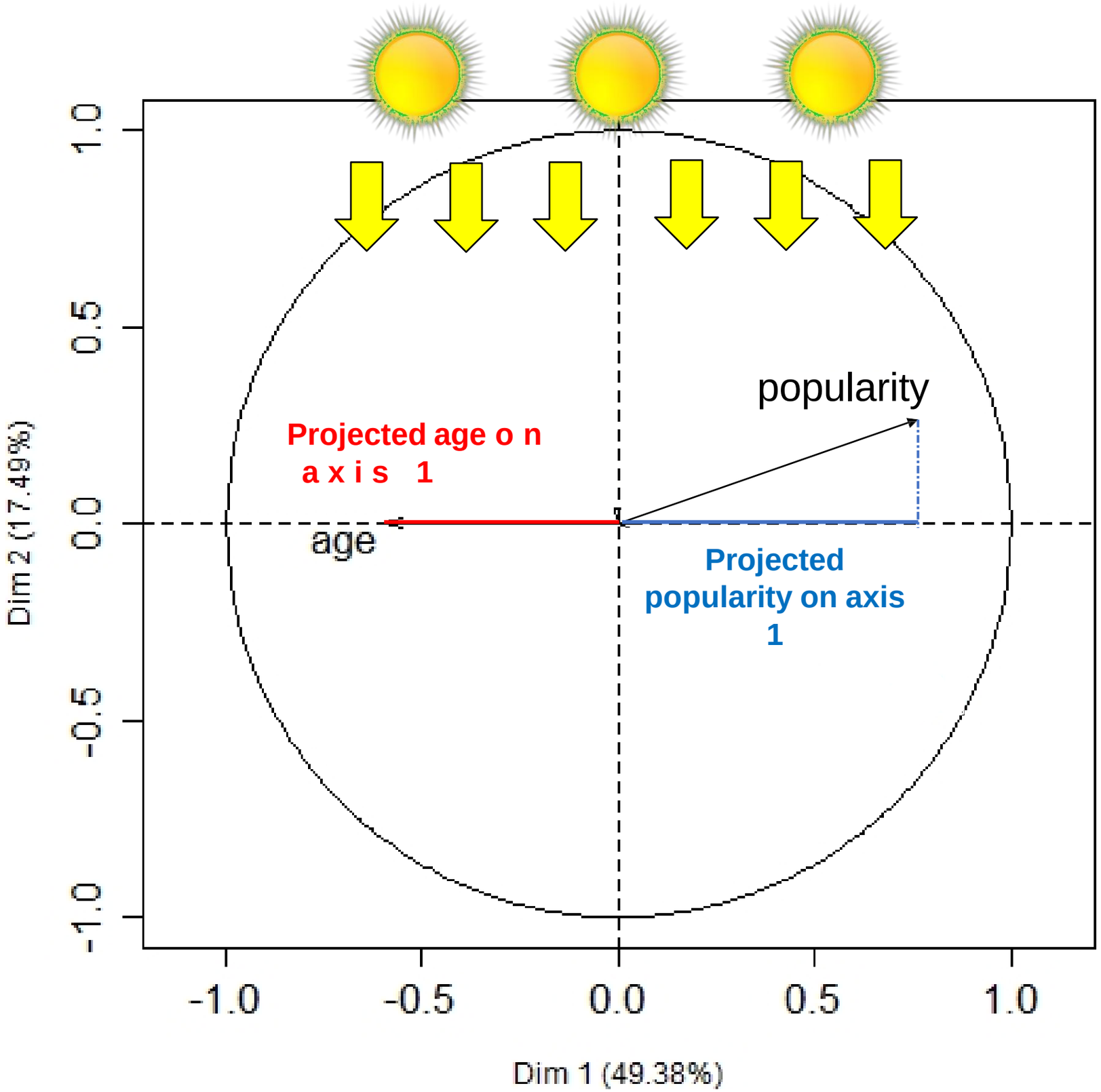
Orthogonal projection of arrows

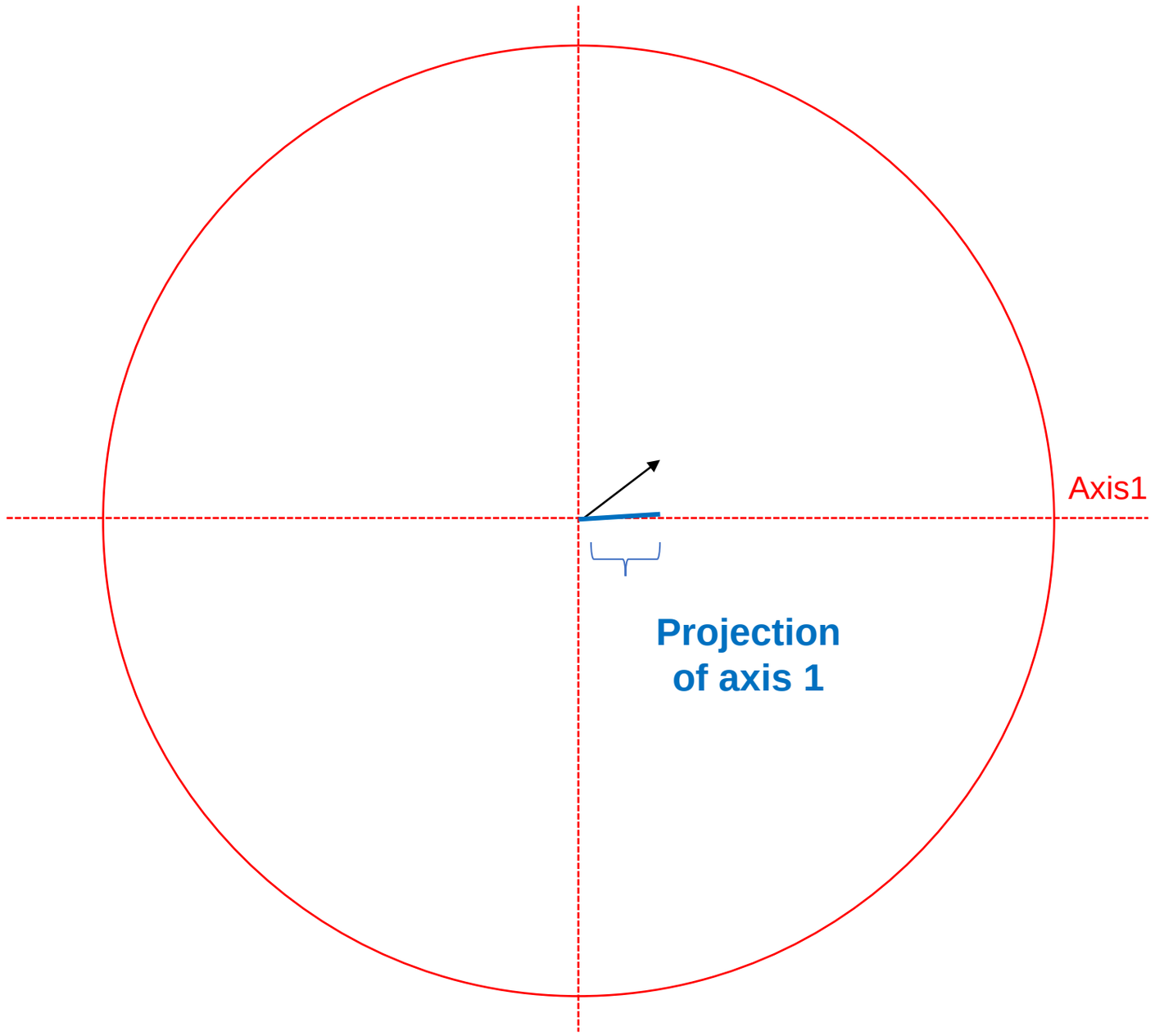


Orthogonal projection of arrows

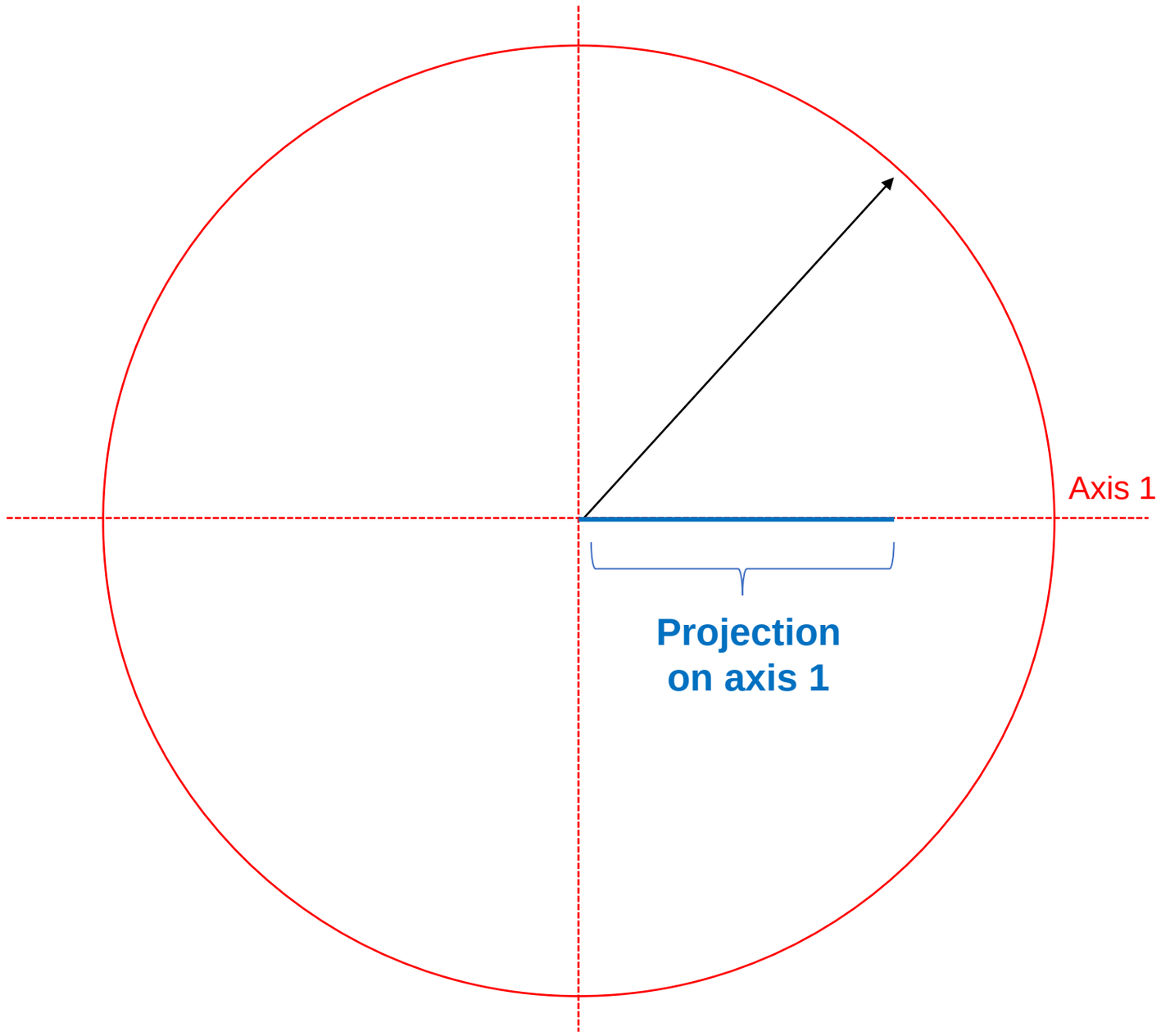

Axis 1
Projection
on axis 1


Graphe des variables de l'ACP
Dim 2 (17.49%)
1.0
0.5
0.0
-0.5
-1.0
Gender
Age
Nb_chp_narr
Popularity
Nb_death_relationships
Married
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (49.38%)

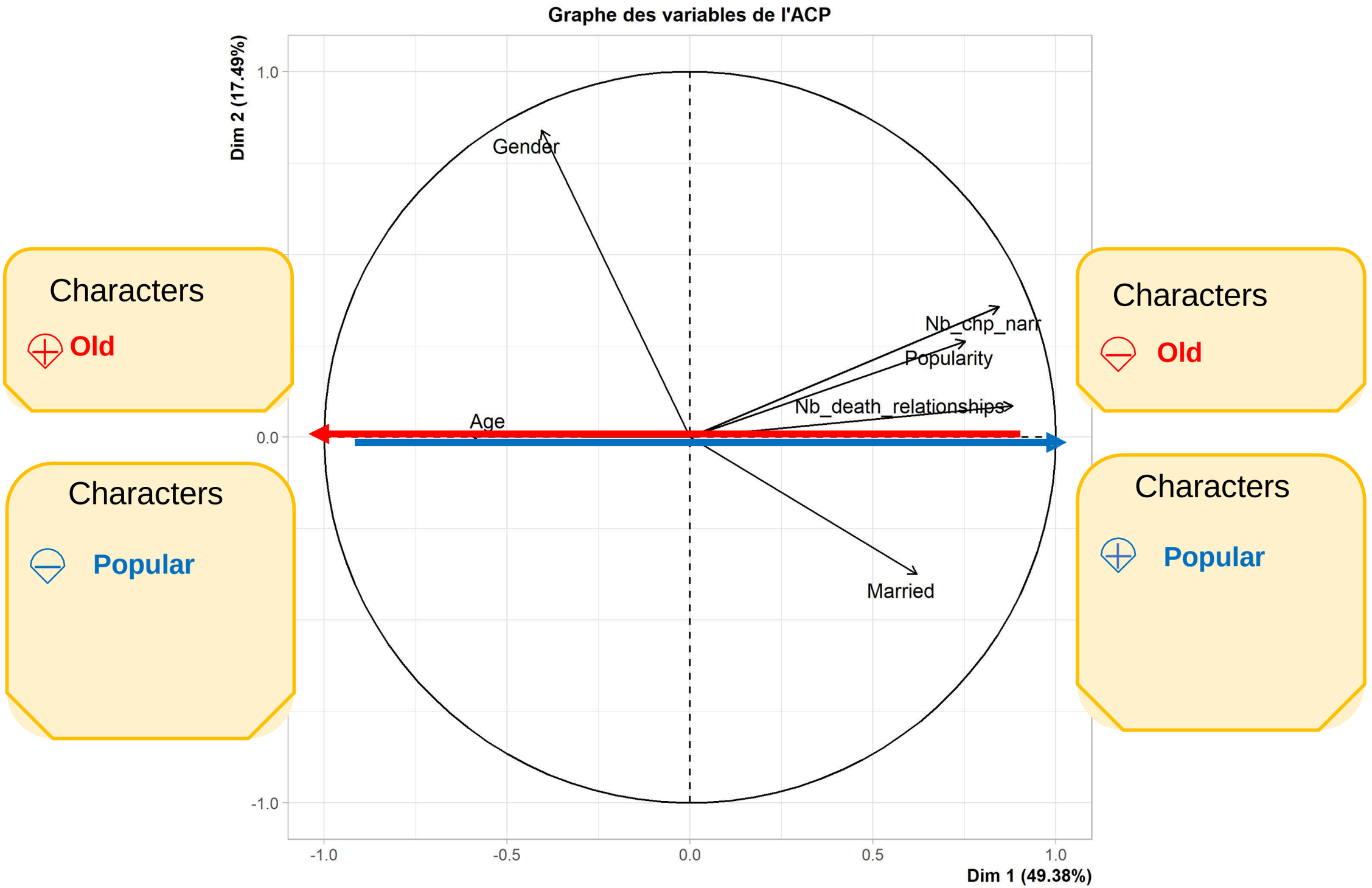

popularity
Projected age o n a x i s 1
age
Projected popularity on axis 1
1.0
0.5
0.0
-0.5
-1.0
Dim 2 (17.49%)
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (49.38%)

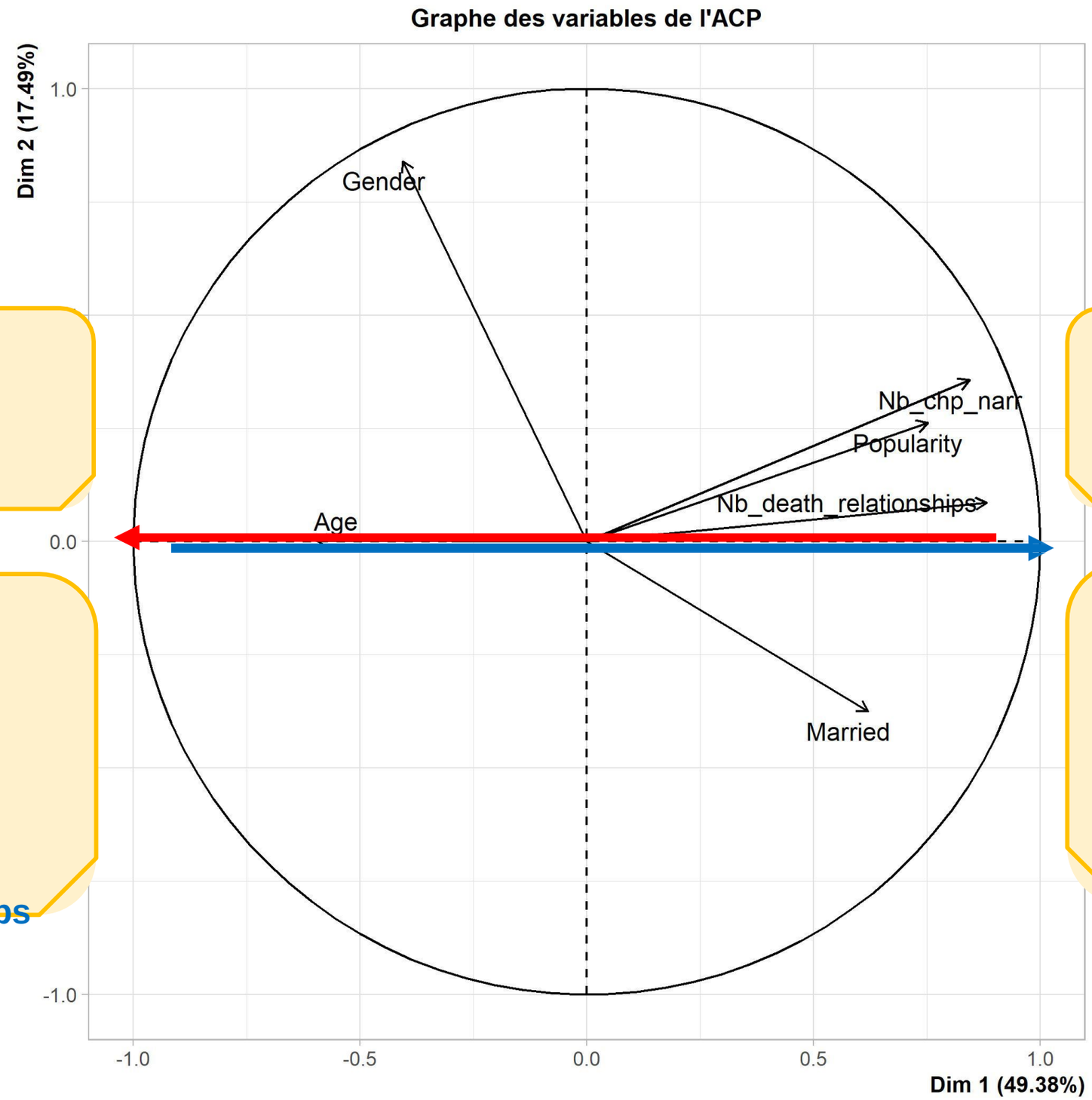


Small projection on axis 1, the variable provides little information on the distribution of points on axis 1.



Larger projection on axis 1, the variable provides more information on the distribution of points on axis 1.


Graphe des variables de l'ACP
Dim 2 (17.49%)
1.0
0.0
-1.0
Gender
Age
Nb_chp_narr
Popularity
Nb_death_relationships
Married
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (49.38%)
Characters
Old
Characters
Old
Characters
Popular
Characters
Popular



Characters

⊕ **Old**

Characters

⊖ **Old**

Characters

⊖ **Popular**

⊖ **Narrators**

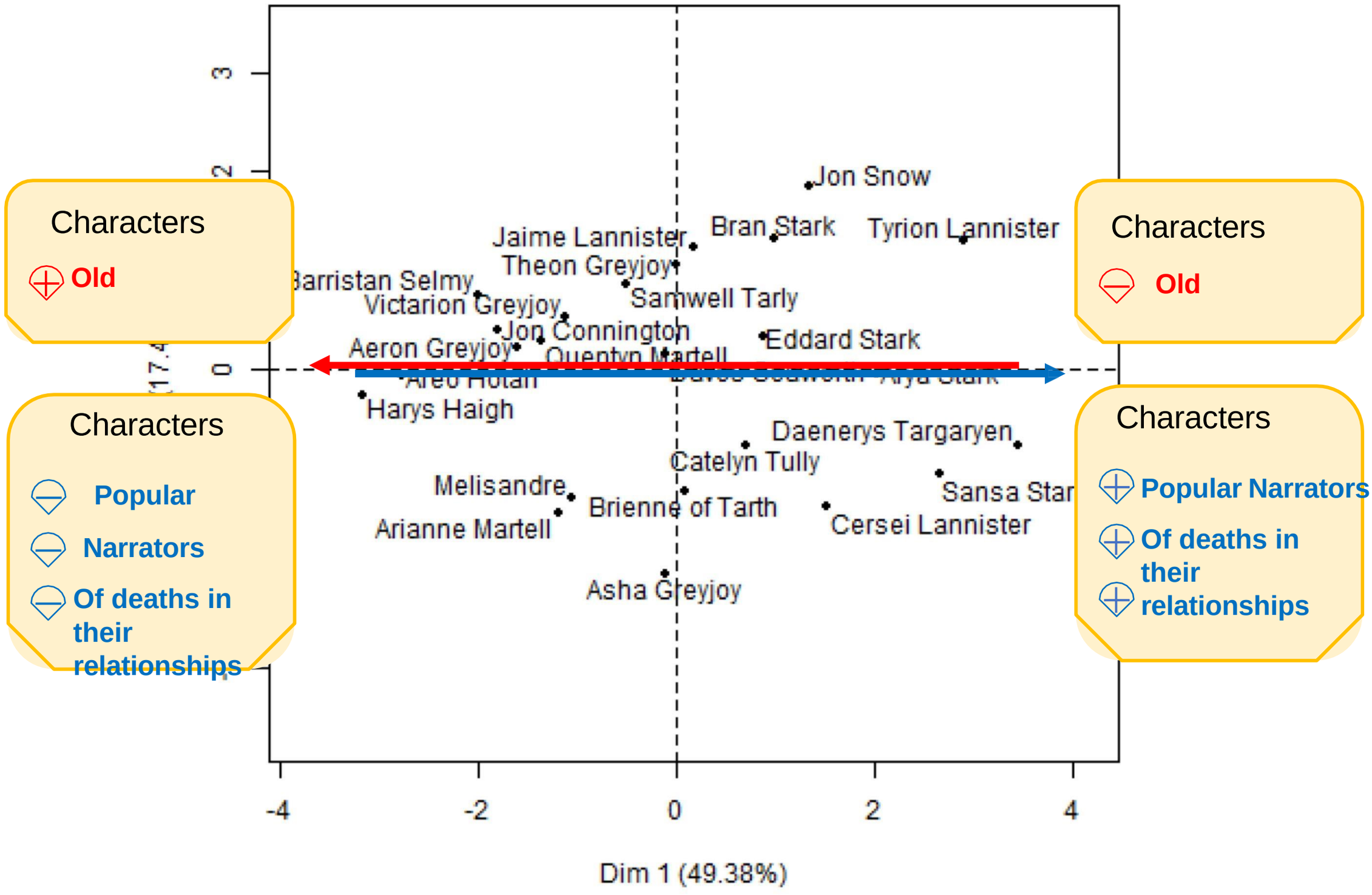
⊖ **Deaths in their relationships**

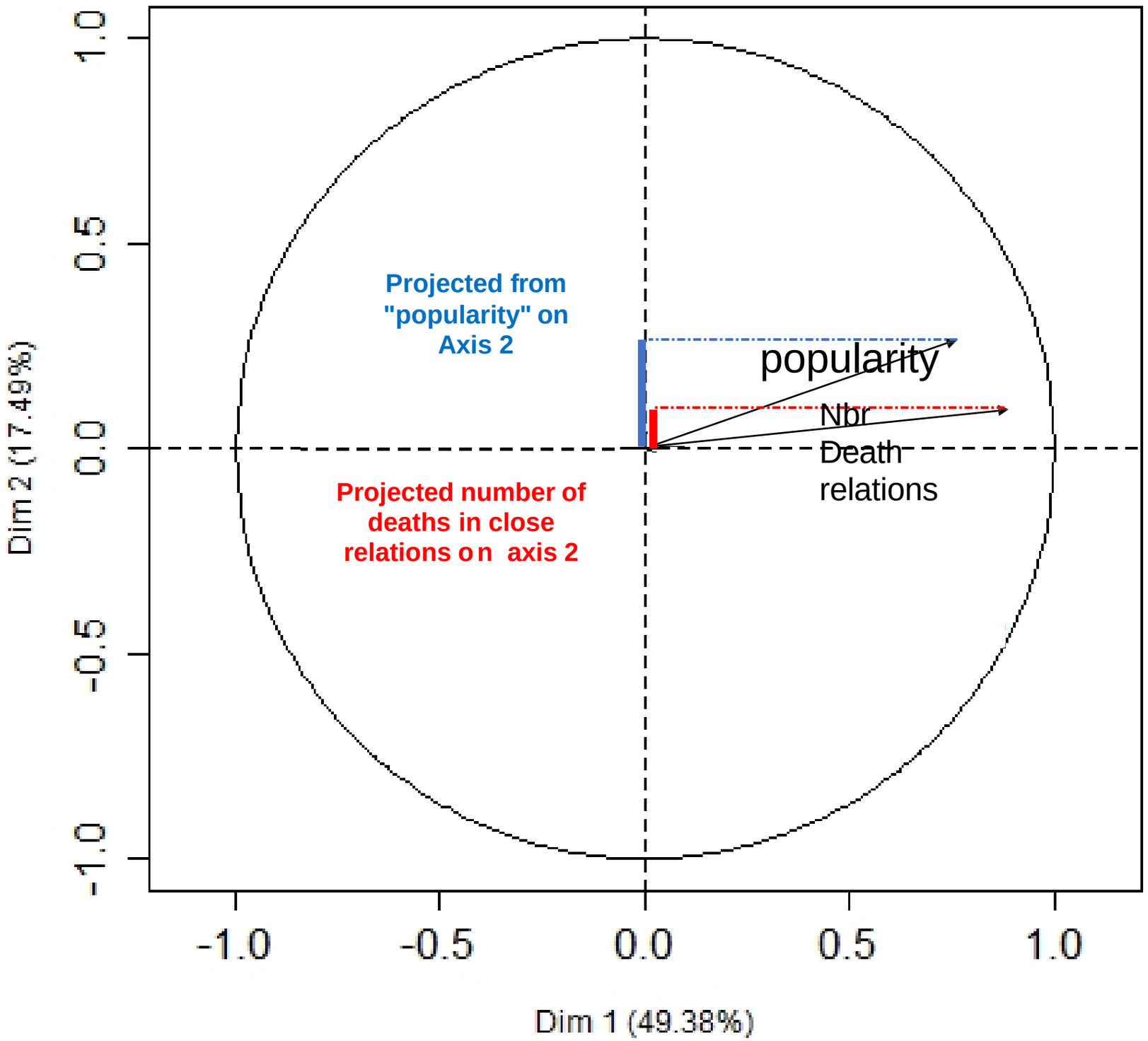
Characters

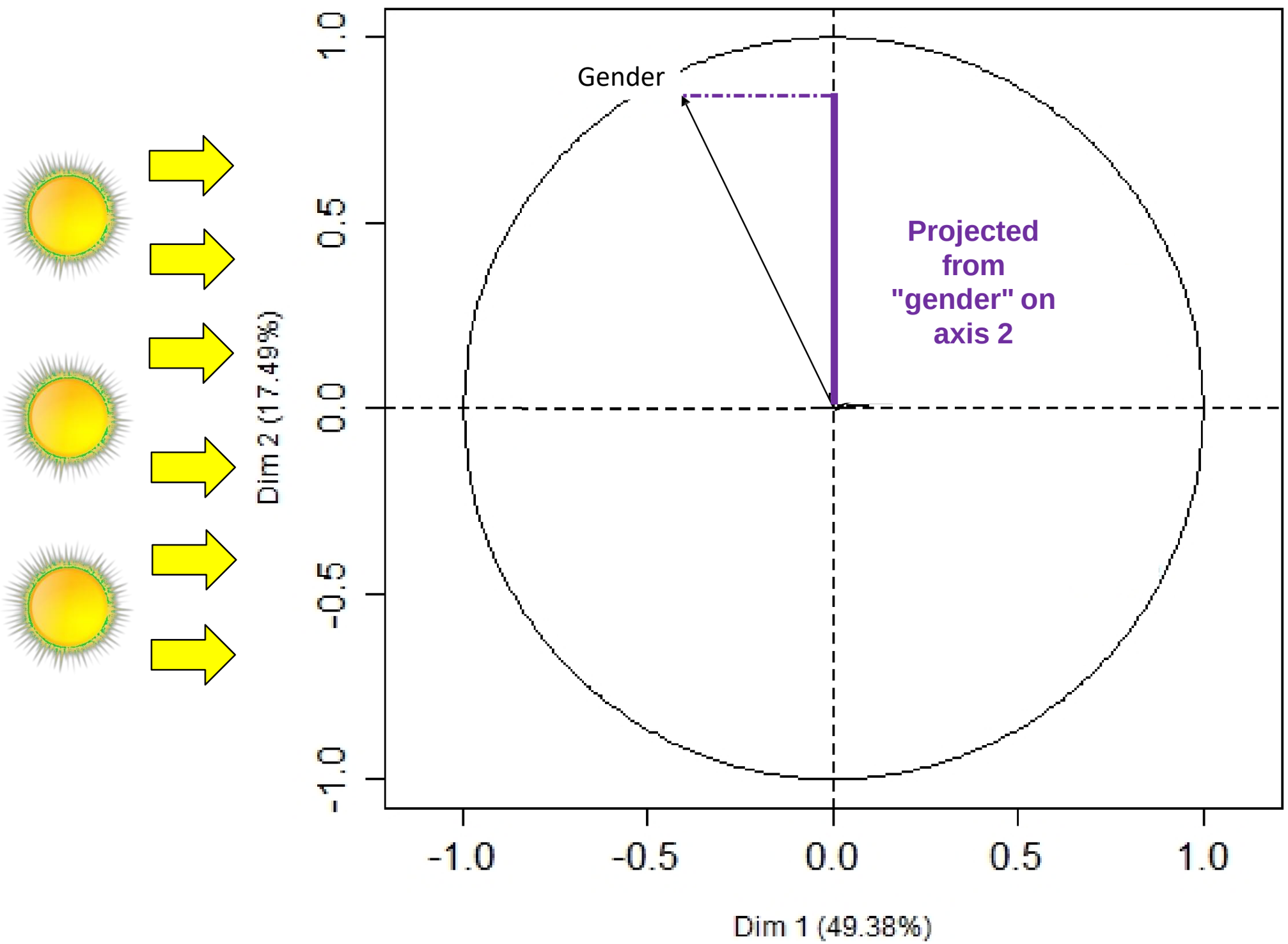
⊕ **Popular**

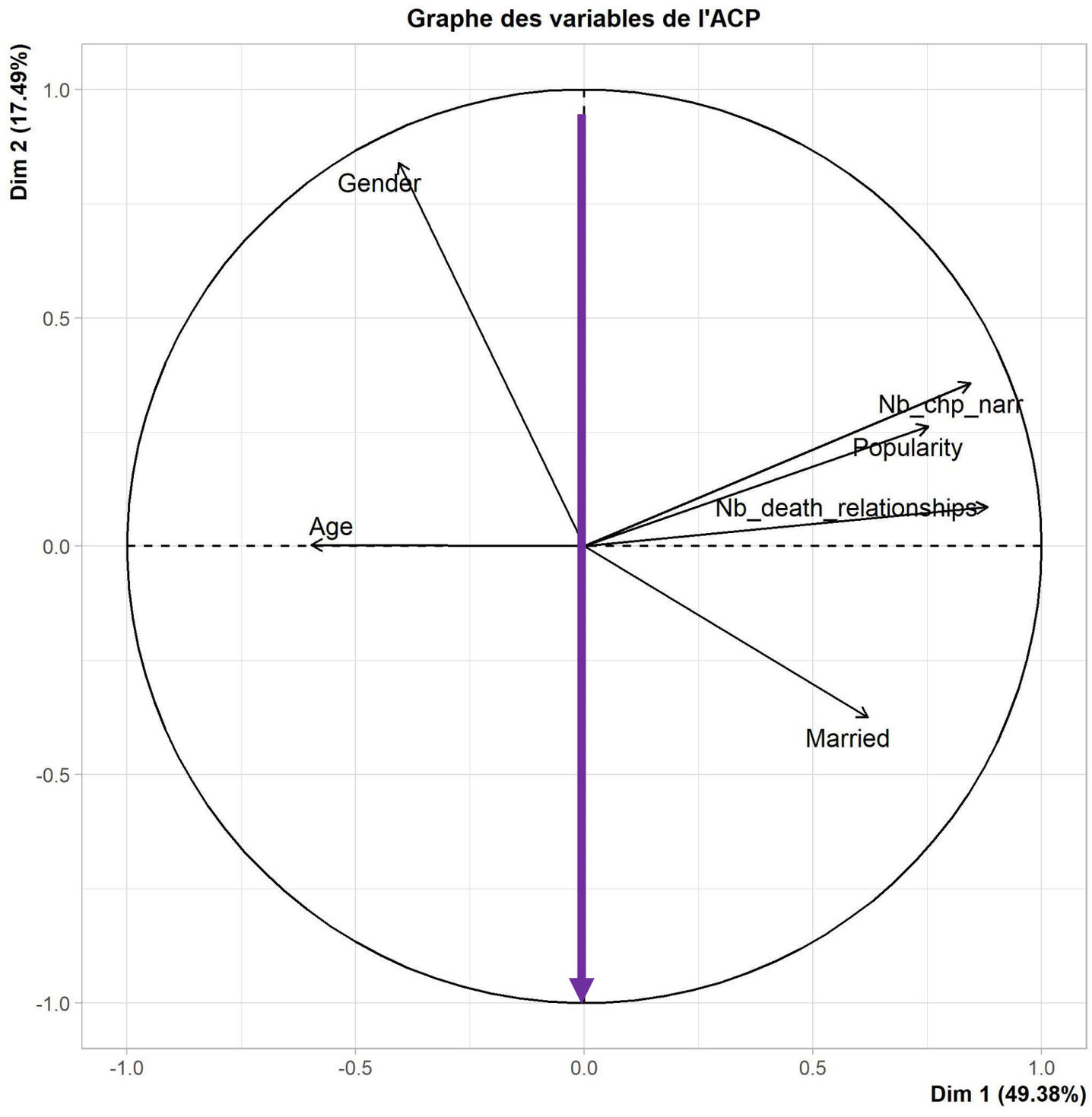
⊕ **Narrators**

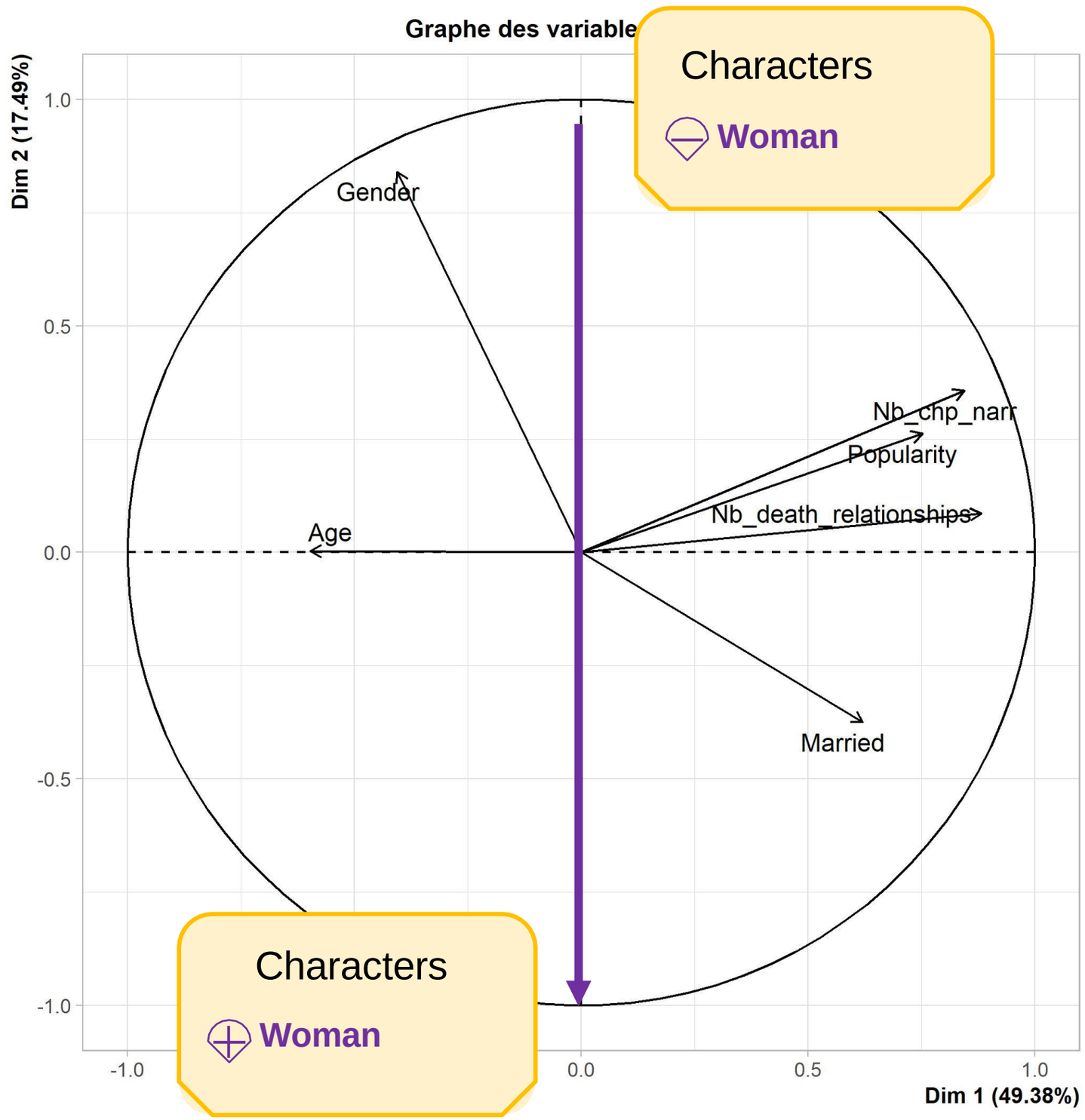
⊕ **Of deaths in their relationships**


Characters
Old
Characters
Old
Characters
Popular
Narrators
Of deaths in their relationships
Characters
Popular Narrators
Of deaths in their relationships
Jon Snow
Bran Stark
Tyrion Lannister
Jaime Lannister
Theon Greyjoy
Samwell Tarly
Victarion Greyjoy
Jon Connington
Eddard Stark
Aeron Greyjoy
Harys Haigh
Daenerys Targaryen
Catelyn Tully
Melisandre
Brienne of Tarth
Sansa Star
Arianne Martell
Cersei Lannister
Asha Greyjoy
3
2
0
-4
-2
0
2
4
Dim 1 (49.38%)

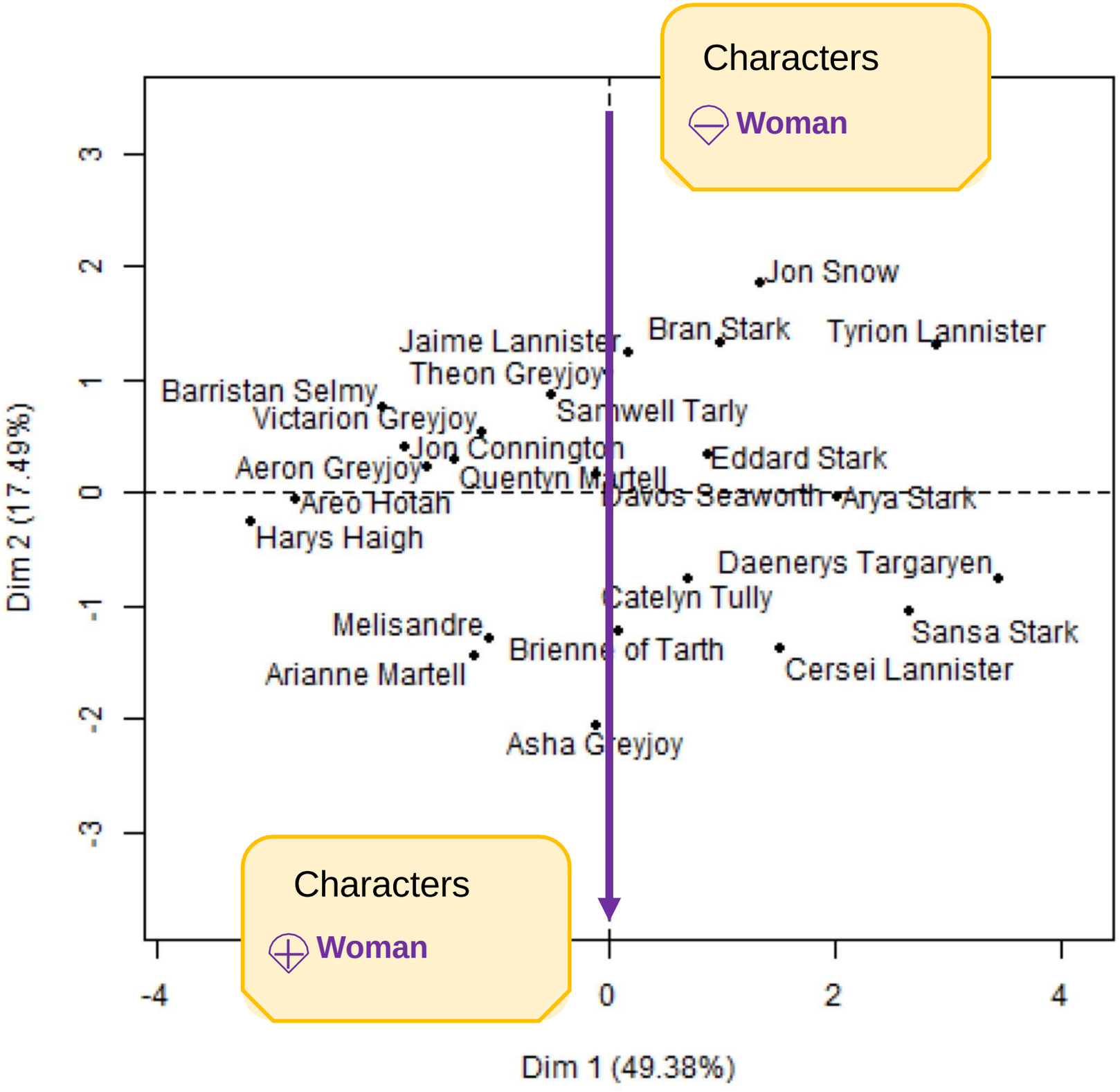

Projected from "popularity" on Axis 2
popularity
Nbr Death relations
Projected number of deaths in close relations on axis 2
1.0
0.5
0.0
-0.5
-1.0
Dim 2 (17.49%)
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (49.38%)

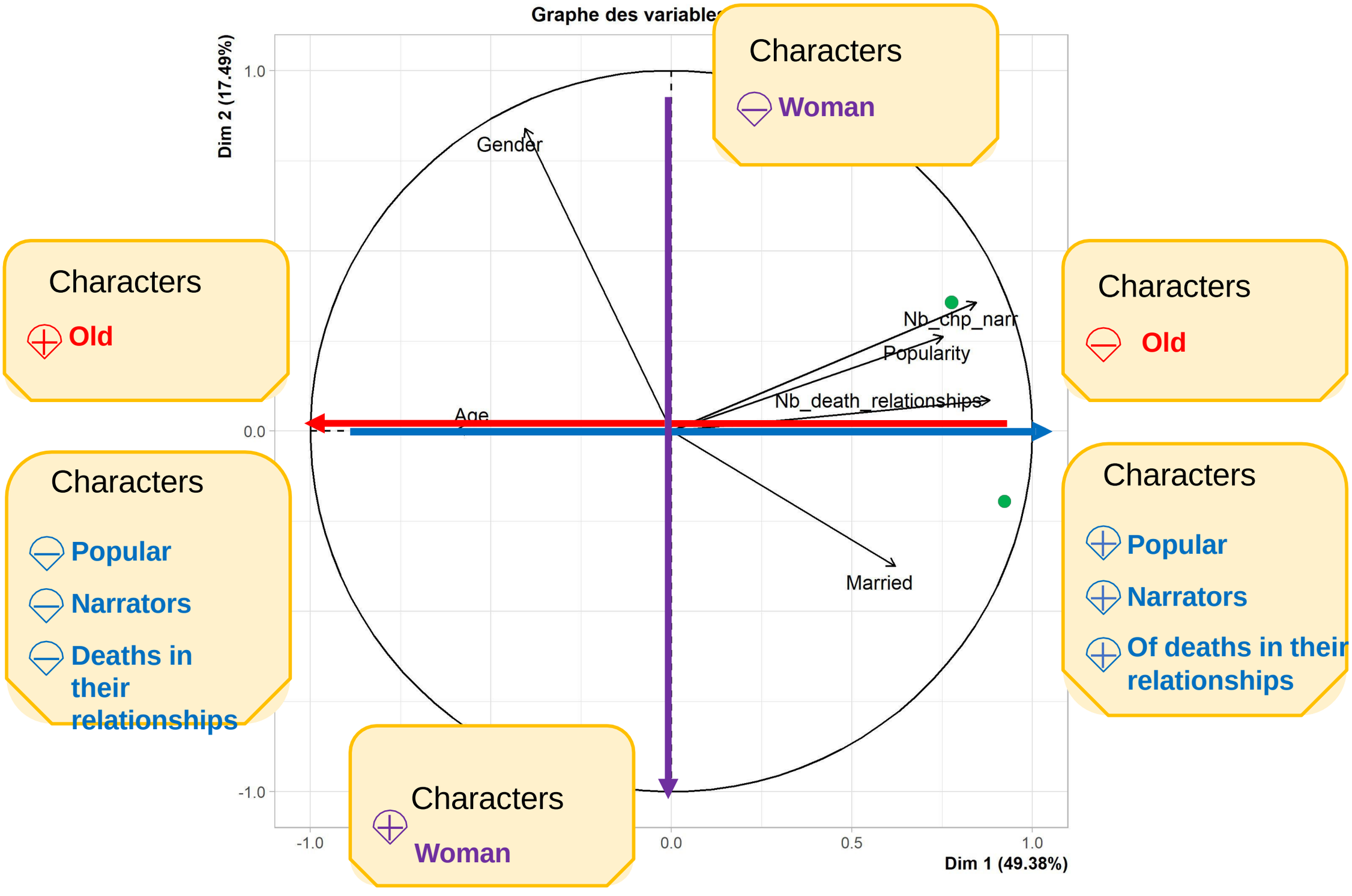

Gender
Projected from "gender" on axis 2
Dim 2 (17.49%)
Dim 1 (49.38%)
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0


Graphe des variables de l'ACP
Dim 2 (17.49%)
1.0
0.5
0.0
-0.5
-1.0
Gender
Age
Nb_chp_narr
Popularity
Nb_death_relationships
Married
-1.0
-0.5
0.0
0.5
1.0
Dim 1 (49.38%)



Gender = 1 : man
Gender = 0 : woman


Characters
Woman
Jon Snow
Bran Stark
Tyrion Lannister
Jaime Lannister
Theon Greyjoy
Barristan Selmy
Victarion Greyjoy
Samwell Tarly
Jon Connington
Eddard Stark
Aeron Greyjoy
Quentyn Martell
Davos Seaworth
Arya Stark
Areo Hotah
Harys Haigh
Daenerys Targaryen
Catelyn Tully
Sansa Stark
Melisandre
Brienne of Tarth
Cersei Lannister
Arianne Martell
Asha Greyjoy
Characters
Woman
3
2
1
0
-1
-2
-3
-4
0
2
4
Dim 2 (17.49%)
Dim 1 (49.38%)


Graphe des variables
Dim 2 (17.49%)
1.0
0.0
-1.0
Gender
Age
Nb_chp_narr
Popularity
Nb_death_relationships
Married
-1.0
0.0
0.5
1.0
Dim 1 (49.38%)
Characters
Woman
Characters
Old
Characters
Old
Characters
Popular
Narrators
Deaths in their relationships
Characters
Woman
Characters
Popular
Narrators
Of deaths in their relationships

Characters

Woman

Characters

Old

Characters

Old

Characters

Popular

Narrators

Of deaths in their relationships

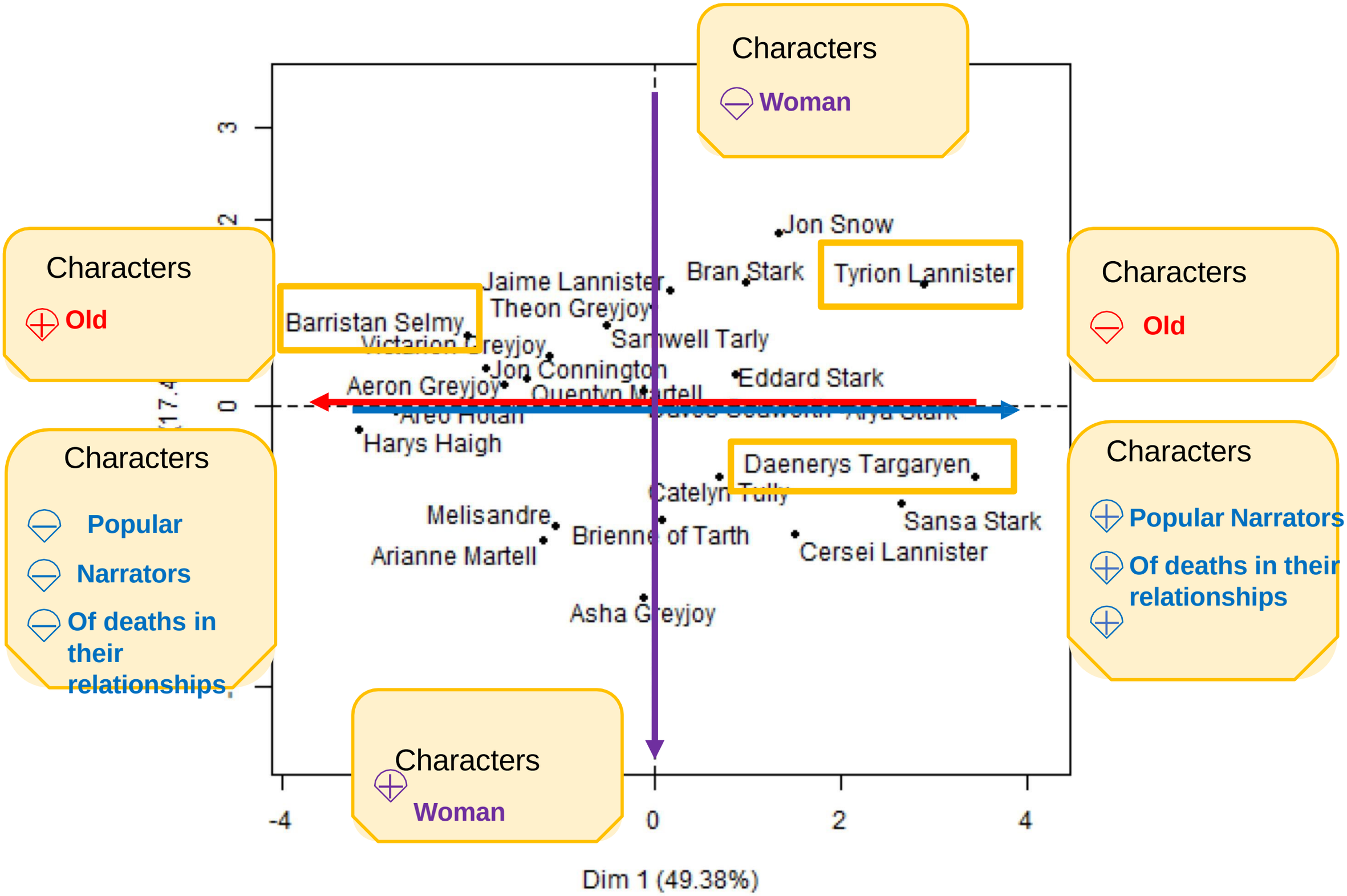
Characters

Popular Narrators

Of deaths in their relationships

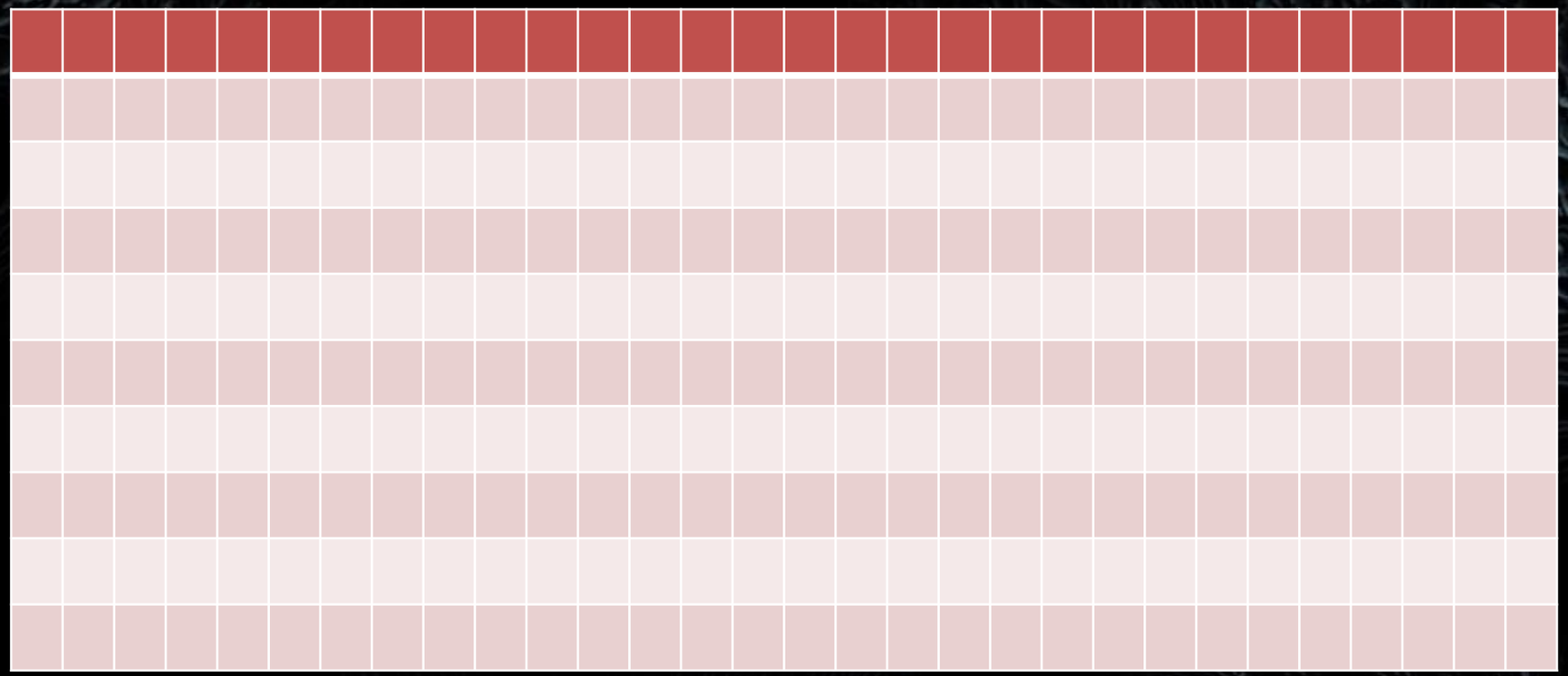
Characters

Woman



Tables wider than they are long

More variables and fewer values

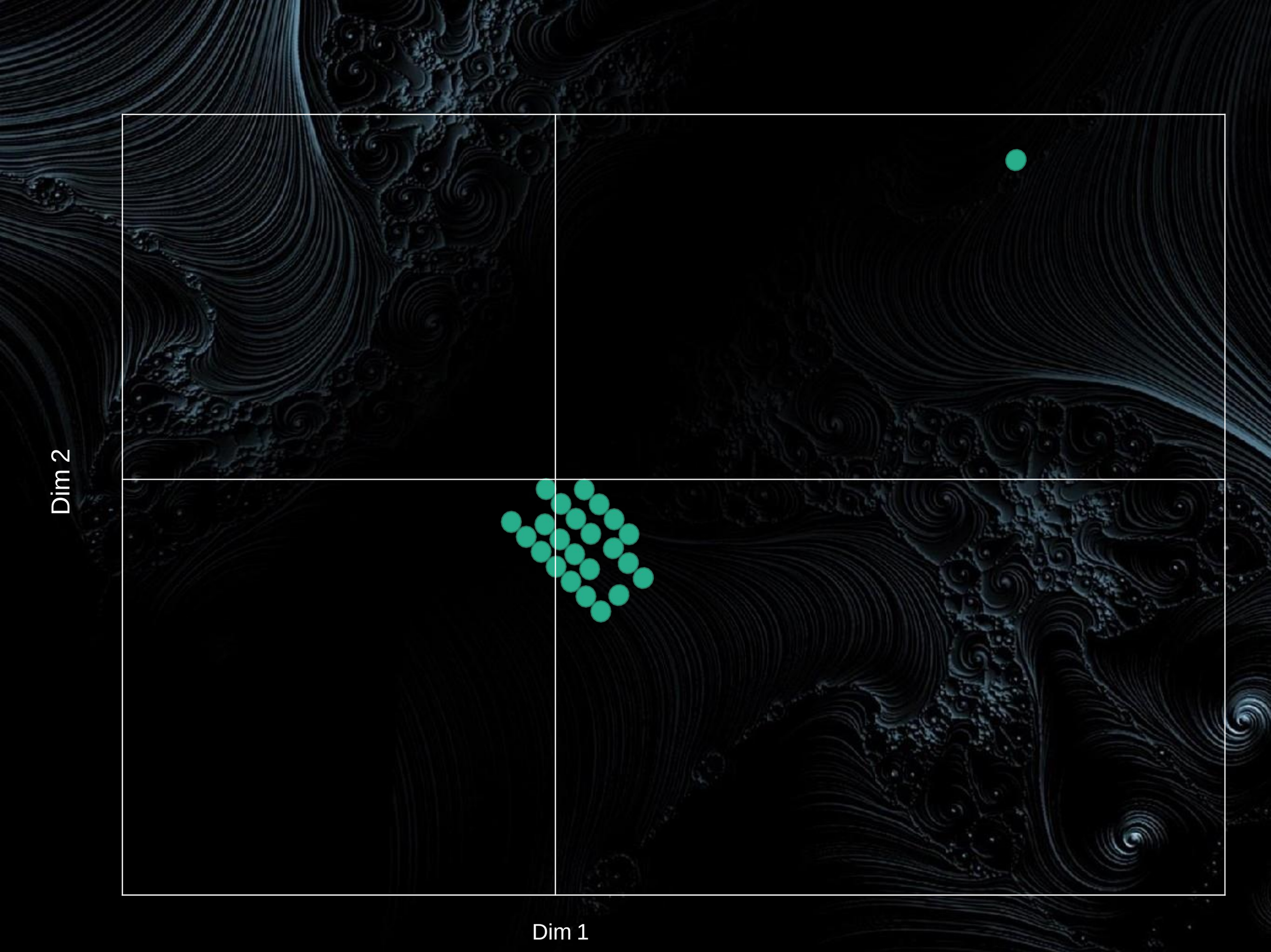


My PCAs show individuals who are very distant from one another, so I really do have different pr[illegible]

Dim 2

Misinterpretation

‘My PCAs show individuals who are very different from one another, so I really do have different profiles.’


Dim 2
Dim 1

Good job ? You can orientate the graph.

- Characterize selected dimension

- « make the difference between right and left»

What do I have on the right ? What do I have on the left?

What kind of characters do I have on the right vs on the left ?

PCA : method for variables reduction

Reduces variable nomber while creating « super/synthetic variables »

We can name those « super variables ».

The same applies to the other dimensions

Sample sentences to write to analyze a PCA

Step 1: I describe inertia percentages

Ex: The first 2 axes of the analysis express **xxx%** of the total inertia of the dataset; this means that xxx% of the total variability of the plot of individuals (or variables) is represented in this plane.

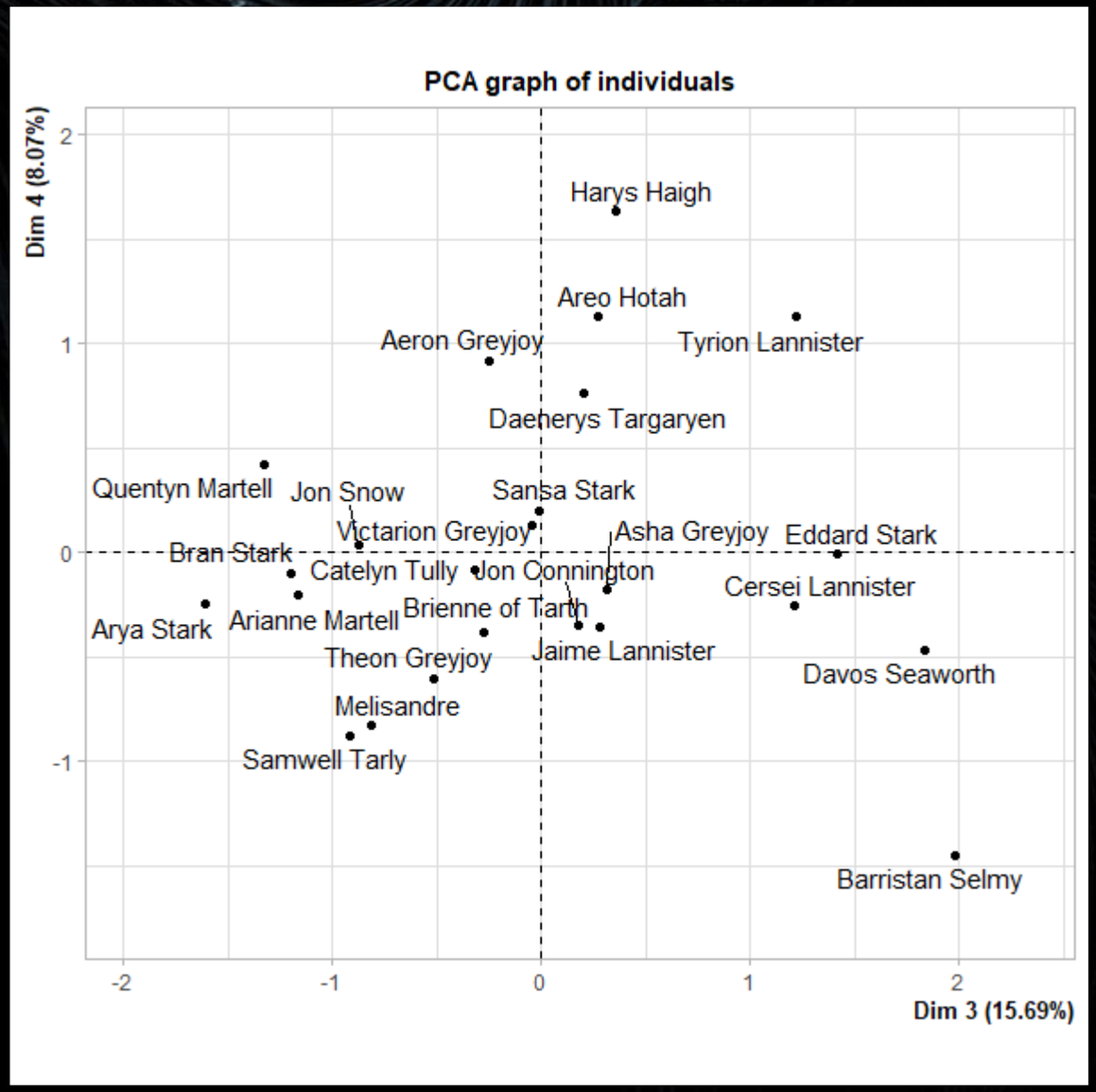
Step 2: I describe the graph of individuals

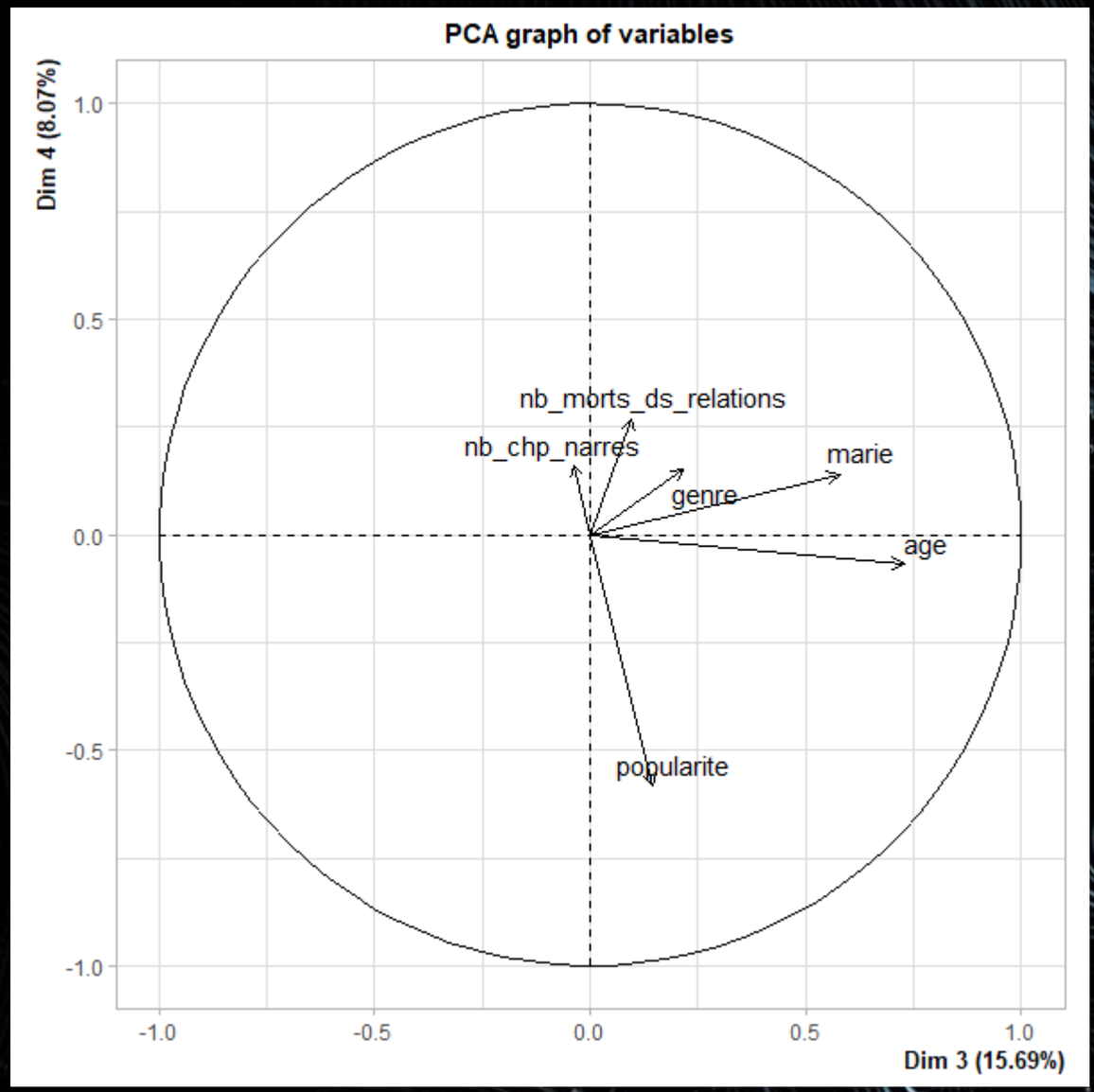
Ex: dimension 1 opposes individuals such as XXX, XXX on the right and individuals such as XXX, XXX on the left.

Step 3: I describe the pot of variables

Ex: Dimension 1 opposes strong XXX values on the left and strong XXX values on the right. I conclude by saying that dimension 1 opposes individuals who tend to be XXX on the right and individuals who tend to be XXX on the left.

Note : we can visualise other dimensions





Dim 3

Dim 4

Dim 20

Dust in higher dimensions

R output interpretation : eigen values

```
> summary(res, nb.dec = 3, nbelements=10, nbind = 10, ncp = 3, file="")

Call:
PCA(X = got.PCA, scale.unit = TRUE, ncp = 5, graph = FALSE)


Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6
Variance               2.963   1.050   0.942   0.484   0.361   0.201
% of var.             49.377  17.493  15.695   8.072   6.016   3.347
Cumulative % of var.  49.377  66.870  82.564  90.637  96.653 100.000
```

Reminder :

% variance ~ % information

R output interpretation: plot of individuals

Distance to points to the barycenter.
We do not care !

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2
Eddard Stark	1.939	0.872	1.069	0.202	0.332	0.437	0.029
Catelyn Tully	1.535	0.712	0.712	0.215	-0.769	2.347	0.251
Sansa Stark	2.884	2.653	9.896	0.846	-1.049	4.364	0.132
Arya Stark	2.746	2.023	5.755	0.543	-0.030	0.004	0.000
Bran Stark	2.134	0.998	1.401	0.219	1.326	6.983	0.386
Jon Snow	2.650	1.342	2.533	0.256	1.843	13.481	0.483
Daenerys Targaryen	3.690	3.457	16.805	0.877	-0.764	2.320	0.043
Tyrion Lannister	3.622	2.895	11.790	0.639	1.294	6.651	0.128
Theon Greyjoy	1.392	0.002	0.000	0.000	1.058	4.448	0.579
Davos Seaworth	2.204	-0.108	0.016	0.002	0.152	0.092	0.005

	Dim.3	ctr	cos2
		8.884	0.534
		0.453	0.043
		0.001	0.000
		11.401	0.342
		6.301	0.313
Jon Snow	-0.875	3.385	0.109
Daenerys Targaryen	0.206	0.187	0.003
Tyrion Lannister	1.220	6.589	0.113
Theon Greyjoy	-0.509	1.147	0.134
Davos Seaworth	1.837	14.927	0.695

Coordinates of individuals on the 1st dimension. > 0 if on the right,
< 0 if on the left

Contribution : does the individual influence a lot axis creation ?

Dim 1

Dim 2

R output interpretation: plot of individuals

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2
Eddard Stark	1.939	0.872	1.069	0.202	0.332	0.437	0.029
Catelyn Tully	1.535	0.712	0.712	0.215	-0.769	2.347	0.251
Sansa Stark	2.884	2.653	9.896	0.846	-1.049	4.364	0.132
Arya Stark	2.746	2.023	5.755	0.543	-0.030	0.004	0.000
Bran Stark	2.134	0.998	1.401	0.219	1.326	6.983	0.386
Jon Snow	2.650	1.342	2.533	0.256	1.843	13.481	0.483
Daenerys Targaryen	3.690	3.457	16.805	0.877	-0.764	2.320	0.043
Tyrion Lannister	3.622	2.895	11.790	0.639	1.294	6.651	0.128
Theon Greyjoy	1.392	0.002	0.000	0.000	1.058	4.448	0.579
Davos Seaworth	2.204	-0.108	0.016	0.002	0.152	0.092	0.005

	Dim.3	ctr	cos2
Eddard Stark	1.417	8.884	0.534
Catelyn Tully	-0.320	0.453	0.043
Sansa Stark	-0.013	0.001	0.000
Arya Stark	-1.605	11.401	0.342
Bran Stark	-1.193	6.301	0.313
Jon Snow	-0.875	3.385	0.109
Daenerys Targaryen	0.206	0.187	0.003
Tyrion Lannister	1.220	6.589	0.113
Theon Greyjoy	-0.509	1.147	0.134
Davos Seaworth	1.837	14.927	0.695

cos2 : is the point well represented ?

R output interpretation: plot of variables

contribution : does the variable contribute a lot to axis creation ?

```
Variables
                          Dim.1    ctr   cos2    Dim.2    ctr   cos2    Dim.3
nb_chp_narres           |  0.845 24.073  0.713 |  0.357 12.127  0.127 | -0.037
genre                   | -0.405  5.544  0.164 |  0.840 67.187  0.705 |  0.217
marie                   |  0.621 13.018  0.386 | -0.375 13.403  0.141 |  0.579
age                     | -0.597 12.022  0.356 |  0.003  0.001  0.000 |  0.727
nb_morts_ds_relations   |  0.882 26.251  0.778 |  0.086  0.711  0.007 |  0.097
popularite              |  0.752 19.092  0.566 |  0.263  6.571  0.069 |  0.144
                           ctr   cos2
nb_chp_narres            0.147  0.001 |
genre                    4.985  0.047 |
marie                   35.539  0.335 |
age                     56.122  0.528 |
nb_morts_ds_relations    0.994  0.009 |
popularite               2.212  0.021 |
```

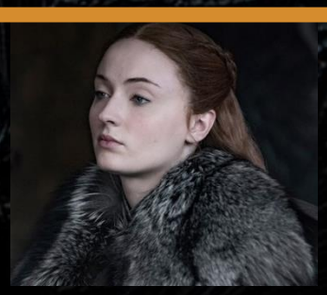
cos2 : is the variable (arrow) well represented on axis ?

Viable contribution

Go there !

Variable which contributes a lot

Dim 1



Please your clients ! Or the reviewers ☺

Supplemental tool : dimdesc

A function from the FactoMineR package listing the most characterisc variables in each dimension in each dimension

```
> dimdesc(res, axes = 1:3, proba = 0.05)
$Dim.1

Link between the variable and the continuous variables (R-square)
=================================================================
                       correlation         p.value
nb_morts_ds_relations    0.8818793 0.000000001234362
nb_chp_narres            0.8445063 0.000000021168780
popularite               0.7520803 0.000022544719220
marie                    0.6210289 0.001201383117250
genre                   -0.4052791 0.049448460055180
age                     -0.5968063 0.002079782566080

$Dim.2

Link between the variable and the continuous variables (R-square)
=================================================================
      correlation         p.value
genre   0.8397395 0.0000002882547

$Dim.3

Link between the variable and the continuous variables (R-square)
=================================================================
      correlation       p.value
age     0.7269764 0.00005723616
marie   0.5785053 0.00306249280
```

Let’s get back to some points of PCA.

When using PCA ?

- "clear the way" kind of tool

- Helps you understand the relationship between variables and individuals

- Choose a limited number of variables (because with the degree of freedom, you can't include everything)

Center-reduce variables before PCA

Centered-reduced value

value

mean

$$x = \frac{X - \mu}{\sigma}$$

Standard deviation

```
scale(data$var)
```

To make the variables comparable, we need to center-reduce them. What does this mean? It means subtracting the mean and dividing by the standard deviation.

- FactoMineR do it by default.

Limits of PCA?

- Quantitative data

2 qualitative variables

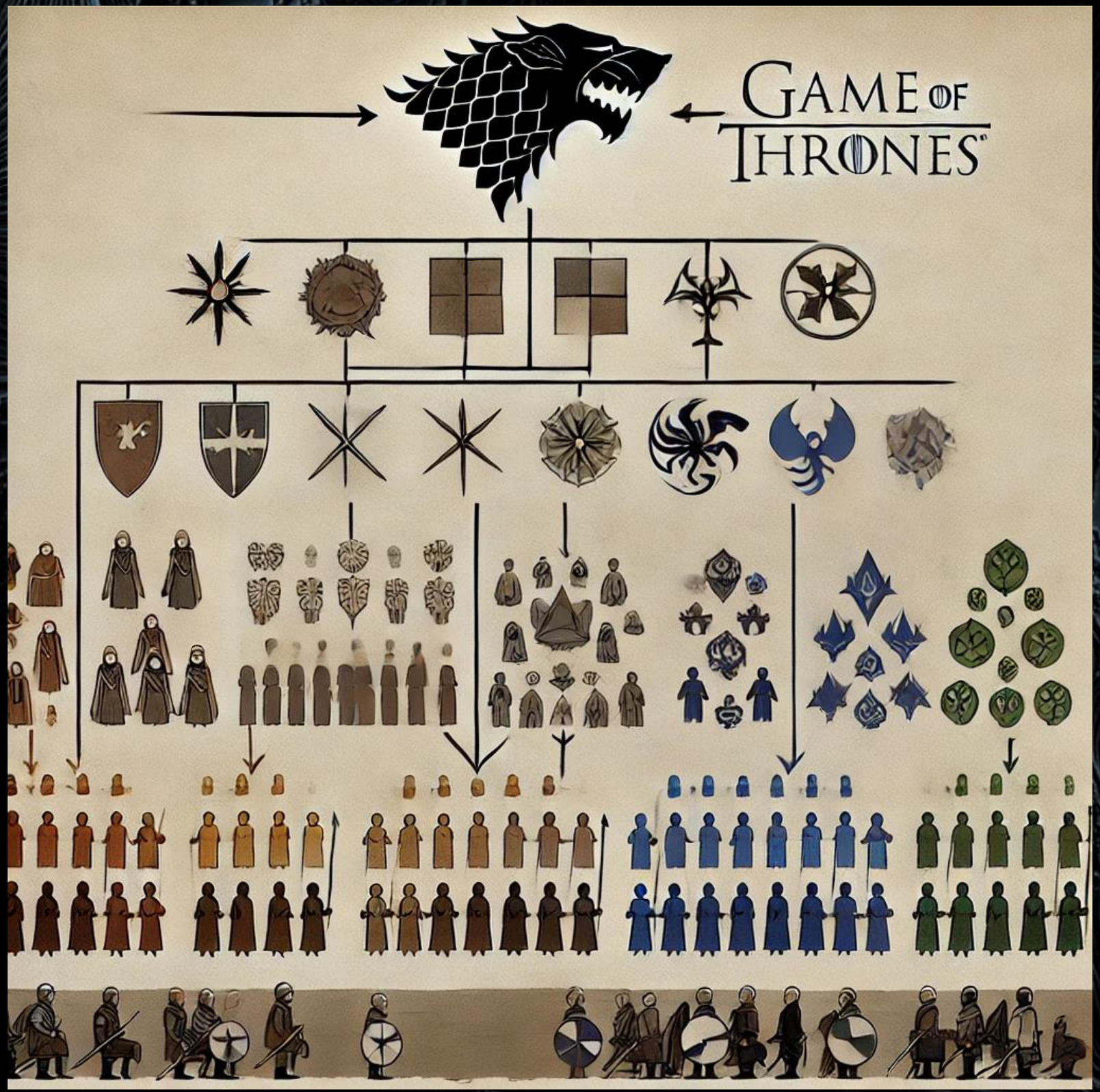
Correspondence Analysis (CA)

> 2 qualitative variables

Multiple Correspondence Analysis (ACM)

Similar = Close on the graph of individuals
Different = Far on the graph of individuals

Classification



Basic idea

Once, we keep the following principle:

- We're close, we're alike

- We're far away, we're different

- => Groups of similar individuals can be created (Group = Cluster)

Customer experience

- Sector analyses can sometimes be a bit hard to get to grips with.

- Clusters speak to everyone.

Cluster

main methods :

- K-means clusters

- Ascending hierarchical clustering (AHC)

The basic principles of these major methods :

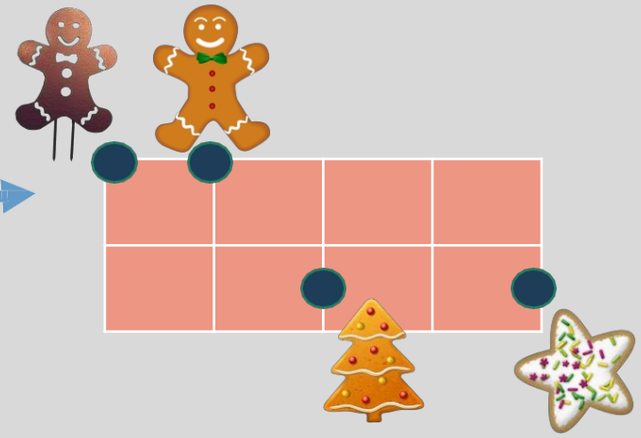
We recover the distances between individuals.

Distance in multidimensional vector space (i.e. as a function of variables).

The closer I am to someone, the more likely I am to be in the same cluster as them.

Building a hierarchical tree

Manhattan's distances



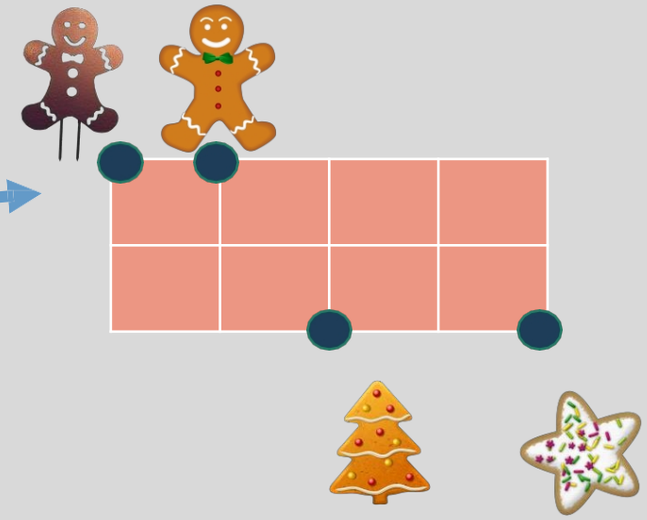
Distance matrix

	0			
	1	0		
	4	3	0	
	6	5	2	0


TAXI
GAME OF THRONES

Building a tree hierarchical

distances from Manhattan



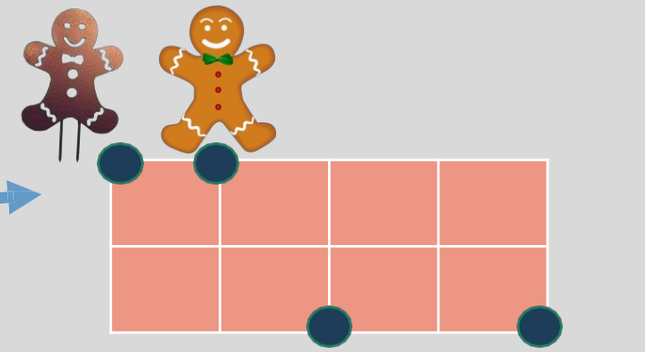
Distance matrix

	0			
	1	0		
	4	3	0	
	6	5	2	0

We are looking for smallest distance.

Building a tree hierarchical

distances from Manhattan



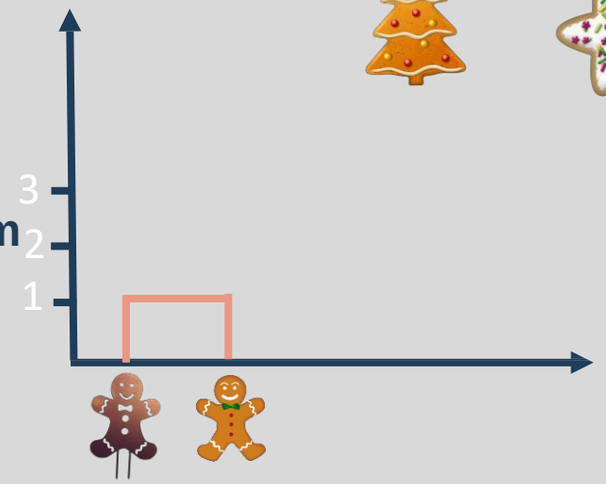
Distance matrix

	0			
	1	0		
	4	3	0	
	6	5	2	0

We are looking for smallest distance.

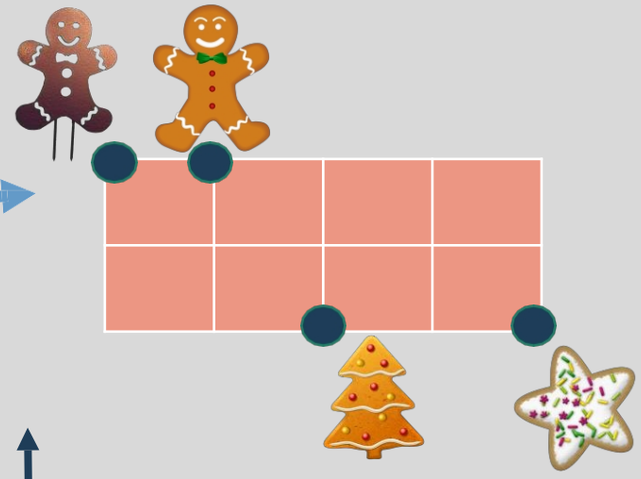
We group

at the height of the distance between them



Building a tree hierarchical

distances from Manhattan



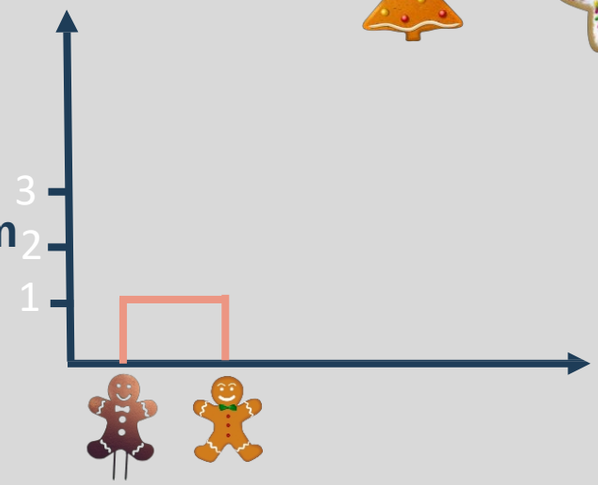
Distance matrix

	0			
	1	0		
	4	3	0	
	6	5	2	0

We are looking for smallest distance.

We group

at the height of the distance between them

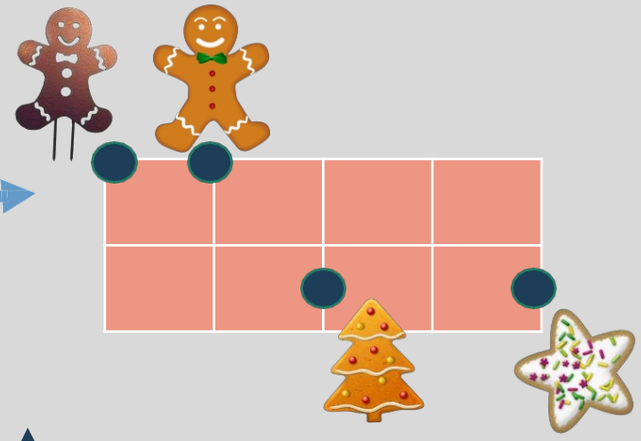


We modify the table with a grouping using the 'minimum jump' criterion: the smallest distance between and

	0		
	3	0	
	5	2	0

Building a tree hierarchical

distances from Manhattan



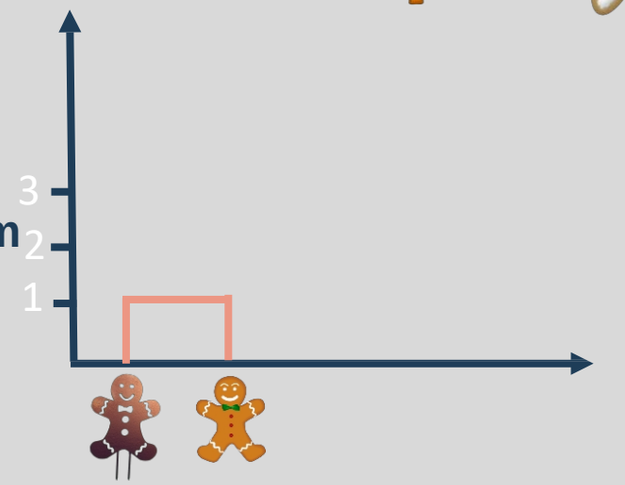
Distance matrix

	0			
	1	0		
	4	3	0	
	6	5	2	0

We are looking for smallest distance.

We group

at the height of the distance between them



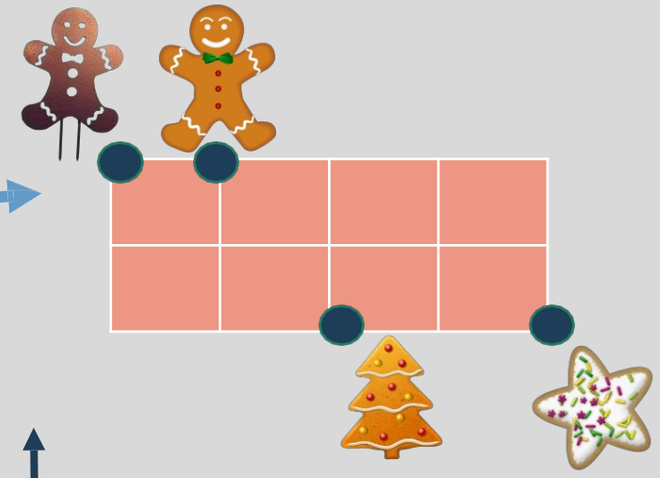
We modify the table with a grouping using the 'minimum jump' criterion: the smallest distance between and

	0		
	3	0	
	5	2	0

Then we look for the smallest distance in the table, and so on.

Building a tree hierarchical

distances from Manhattan



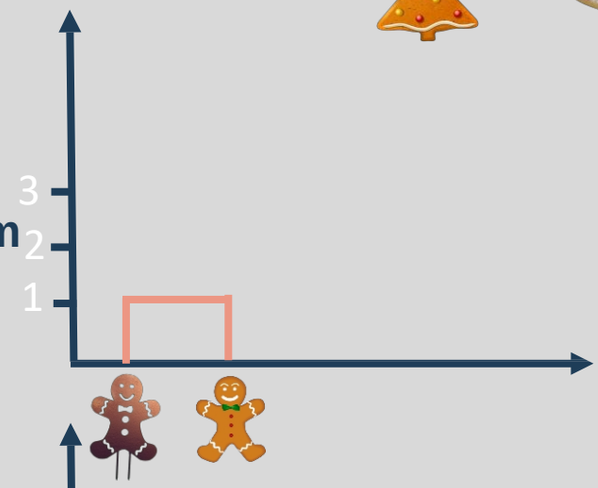
Distance matrix

	0			
	1	0		
	4	3	0	
	6	5	2	0

We are looking for smallest distance.

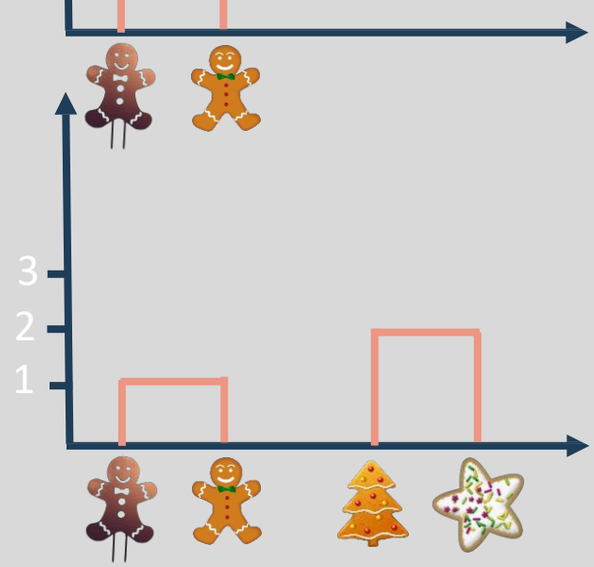
We group

at the height of the distance between them



Modify the table with a grouping using the 'jump' criterion minimum': the most small distance between and

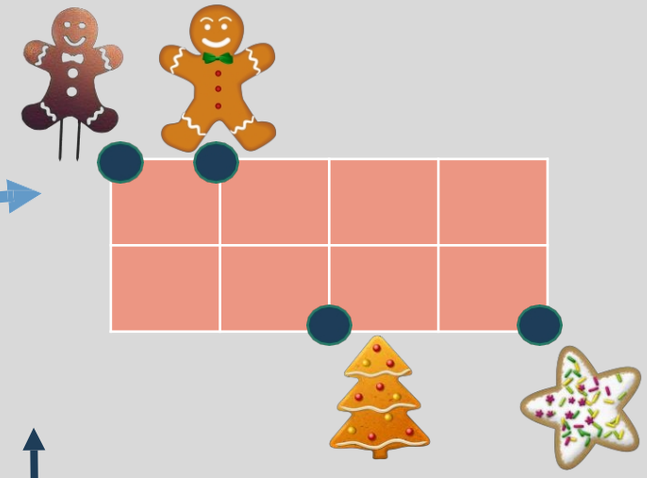
	0		
	3	0	
	5	2	0



Then we look for the smallest distance in the table, and so on.

Building a tree hierarchical

distances from Manhattan



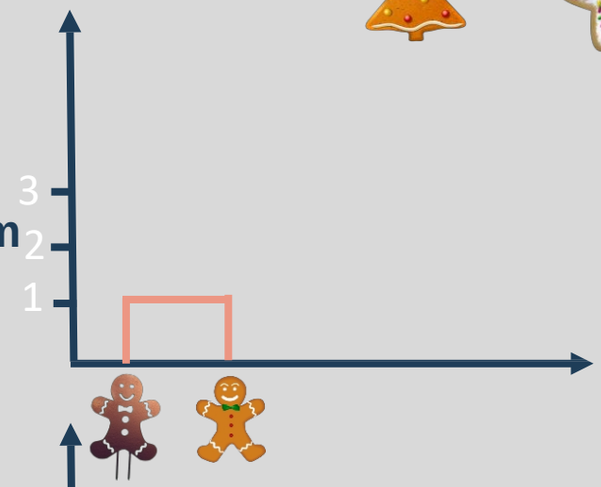
Distance matrix

	0			
	1	0		
	4	3	0	
	6	5	2	0

We are looking for smallest distance.

We group

at the height of the distance between them

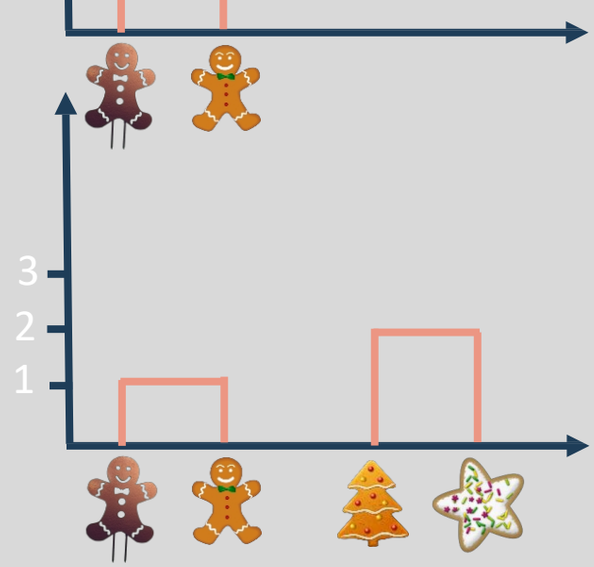


Modify the table with a grouping using the 'jump' criterion minimum': the most small distance between and

	0		
	3	0	
	5	2	0

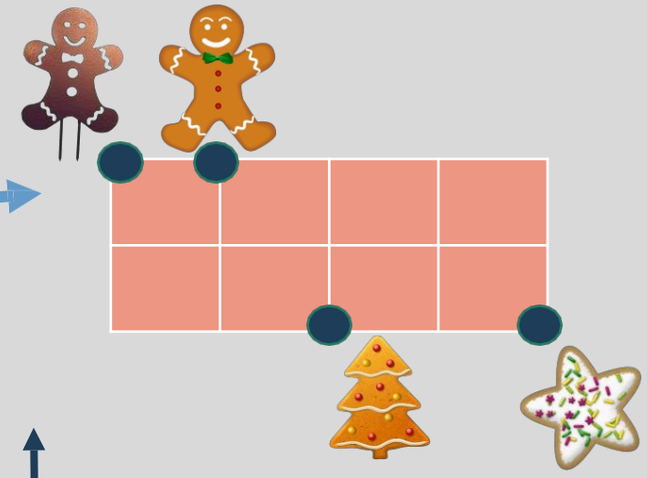
Then we look for the smallest distance in the table, and so on.

	0	
	3	0



Building a tree hierarchical

distances from Manhattan



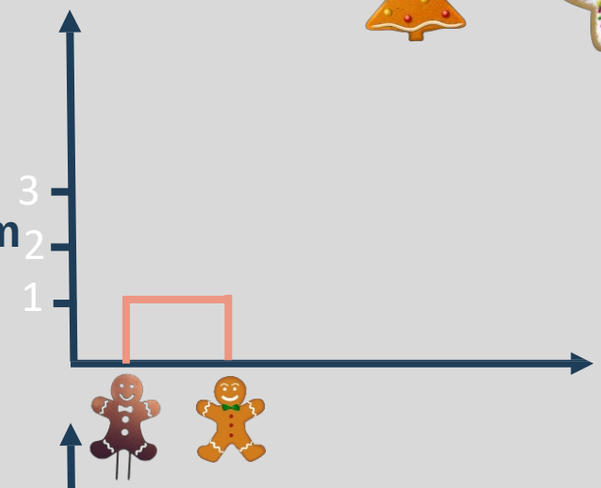
Distance matrix

	0			
	1	0		
	4	3	0	
	6	5	2	0

We are looking for smallest distance.

We group

at the height of the distance between them

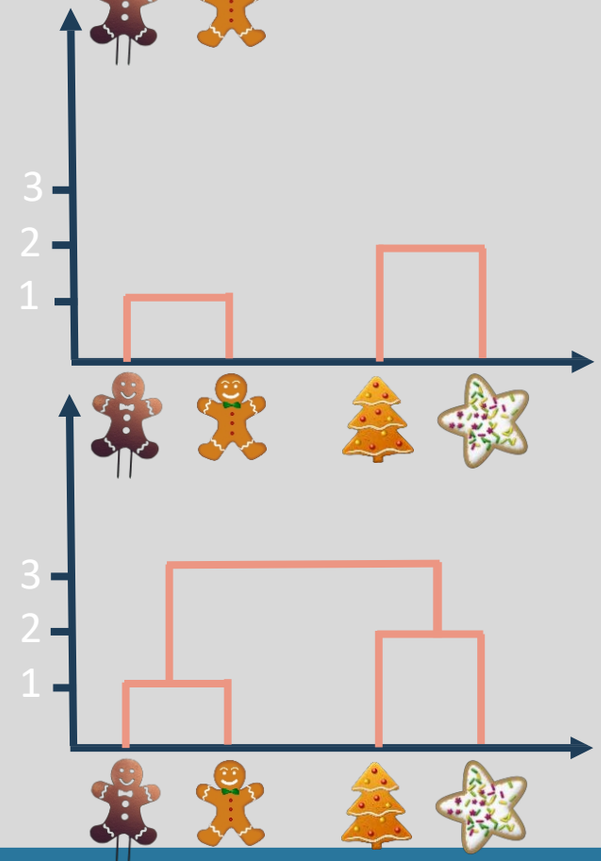


Modify the table with a grouping using the 'jump' criterion minimum': the most small distance between and

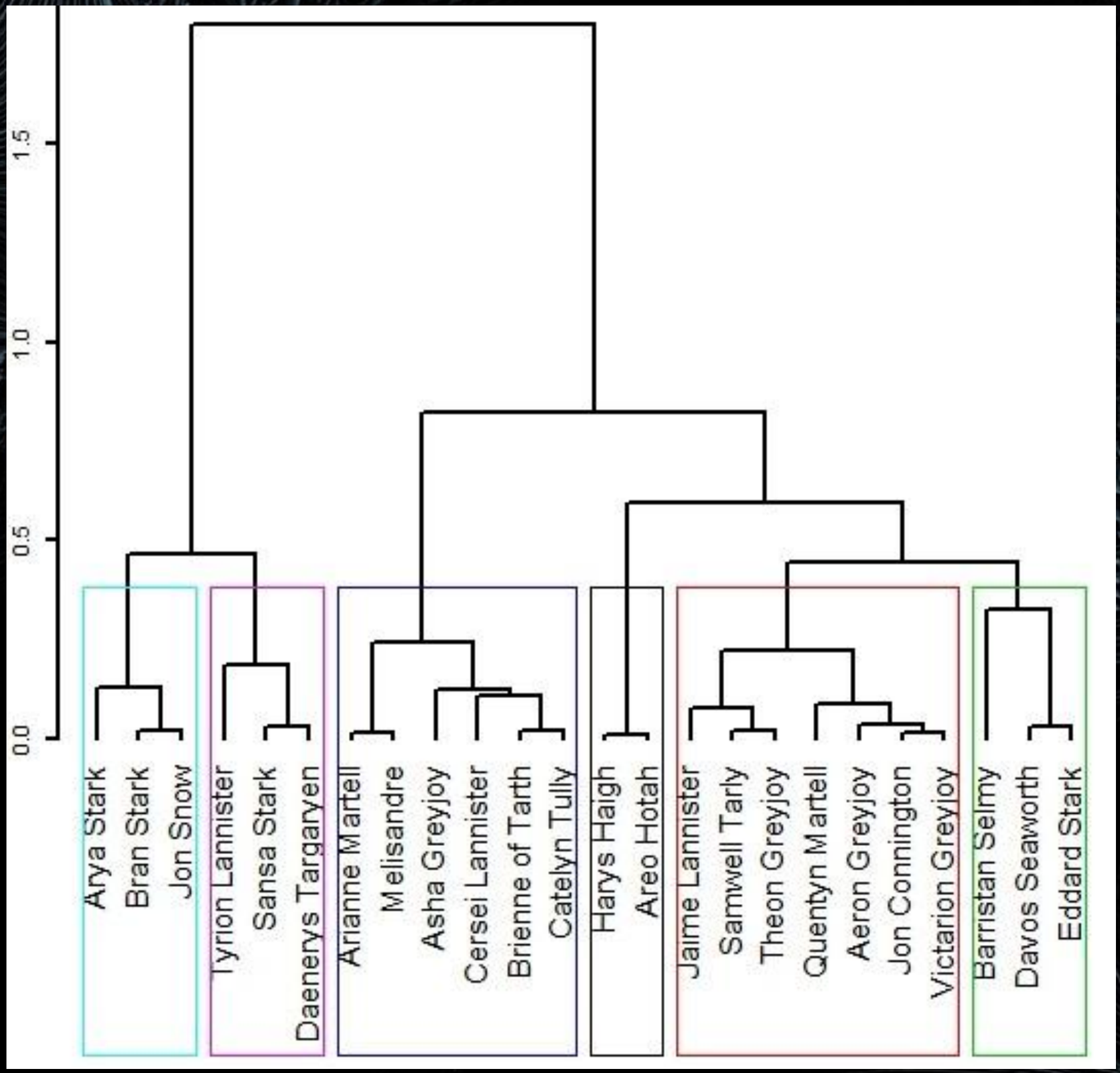
	0		
	3	0	
	5	2	0

Then we look for the smallest distance in the table and so on. continued.

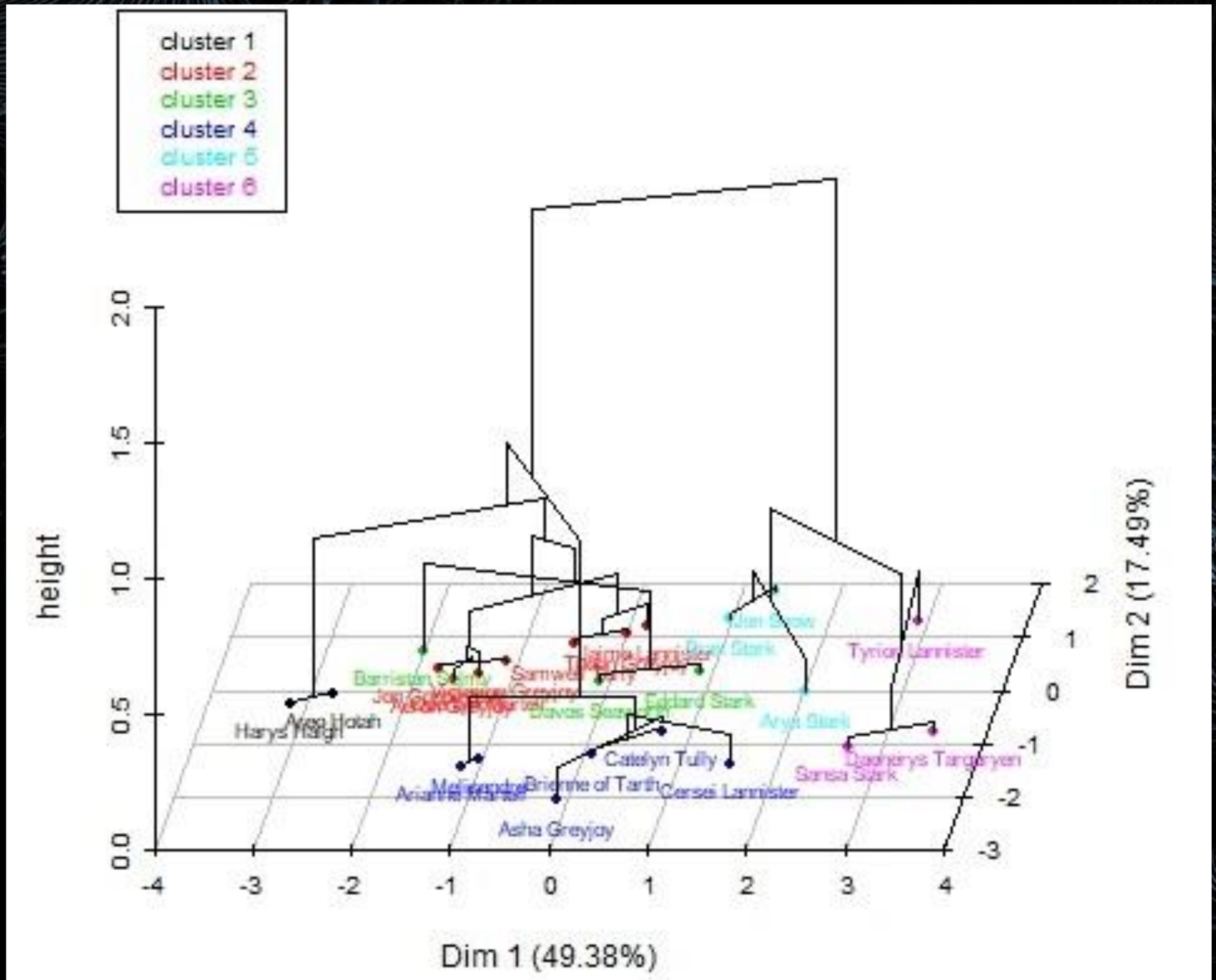
	0	
	3	0



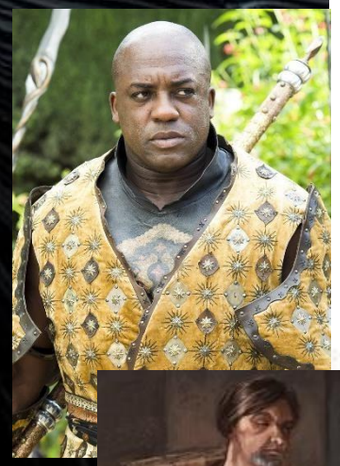
Clustering: hierarchical classification

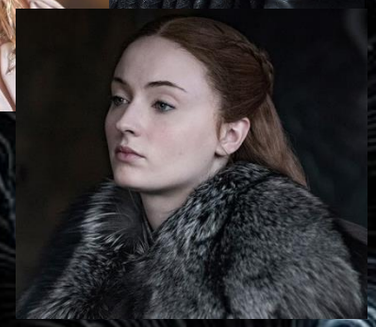


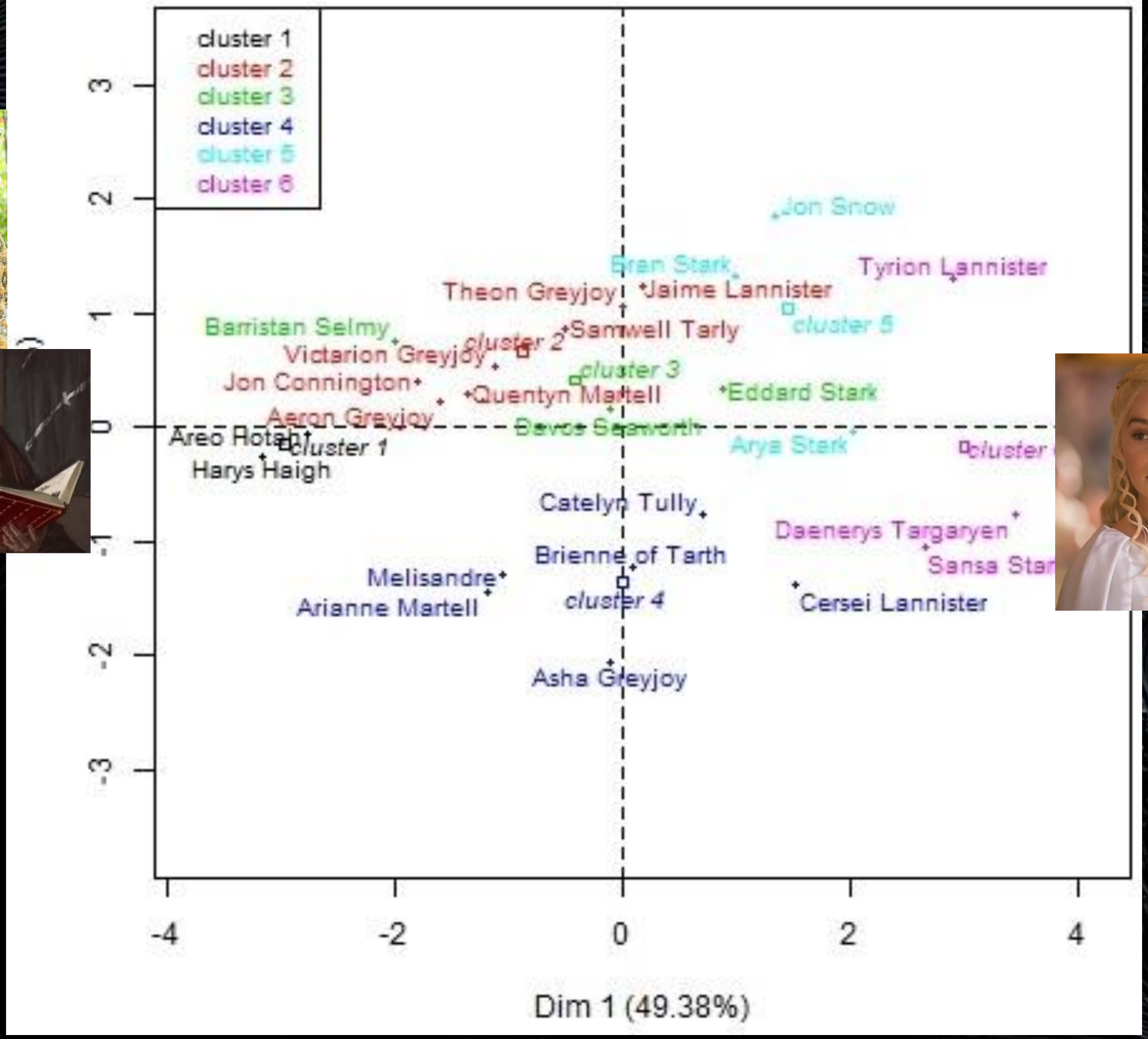
Distance display



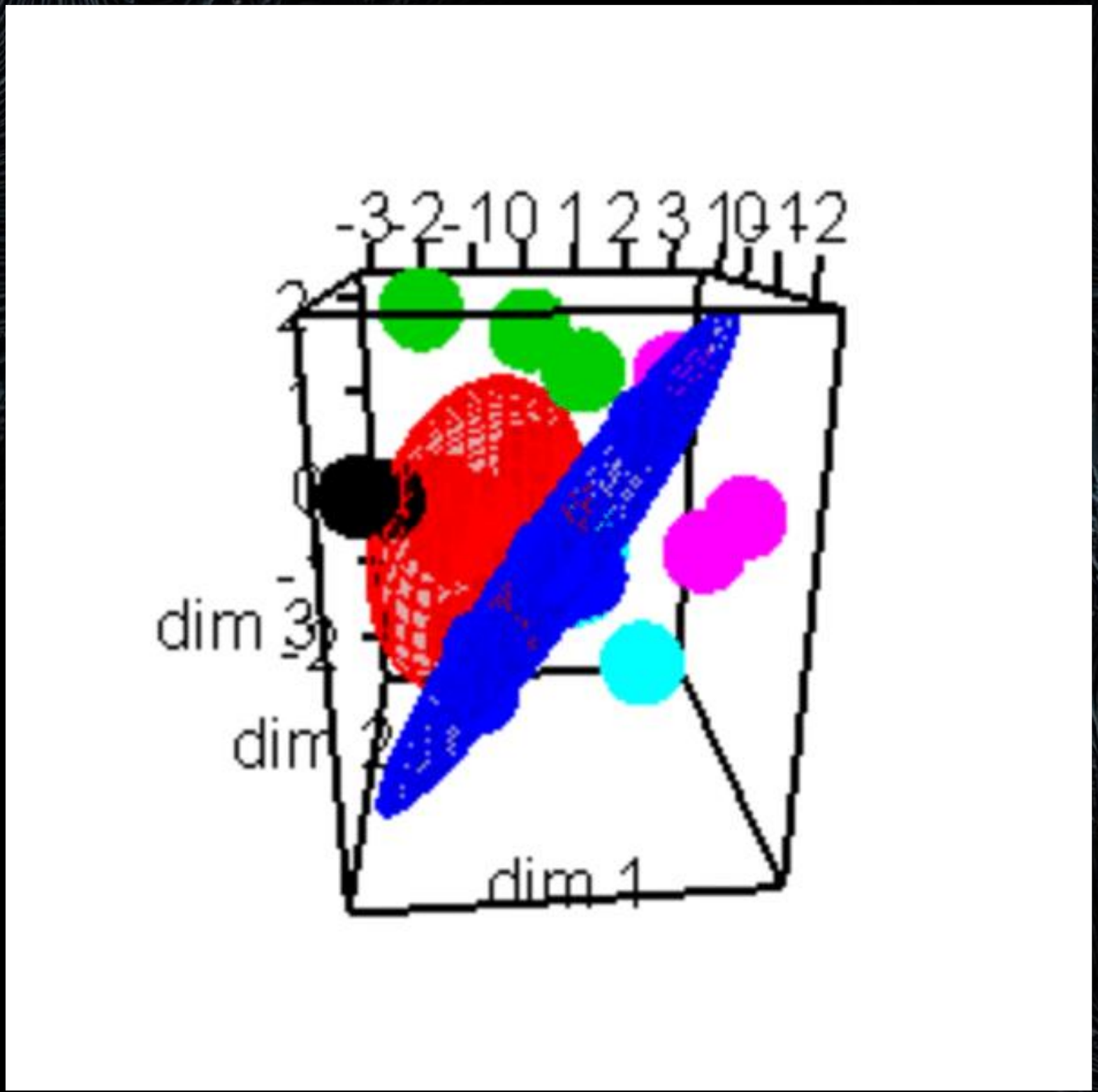
Clustering: hierarchical classification







3D visualisation



K-means algorithm

Fast clustering algorithm.

We specify the number K of clusters, in this case 2.

Steps 2 and 3 are repeated in a loop until the centers are no longer
no longer move (stabilization).

2 disadvantages :
- we choose the number of classes K a priori.
- partition depends on initialization

K-means algorithm

Fast clustering algorithm.

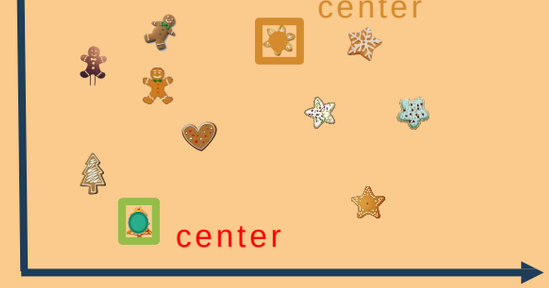
We specify the number K of clusters, in this case 2.

Steps 2 and 3 are repeated until the centers no longer move (stabilization).

2 disadvantages :

- we choose the number of classes K a priori.
- the partition depends on initialization

initialization



Step 1: Randomly designate 2 initial centers in the data, in this case 2 cookies.

K-means algorithm

Fast clustering algorithm.

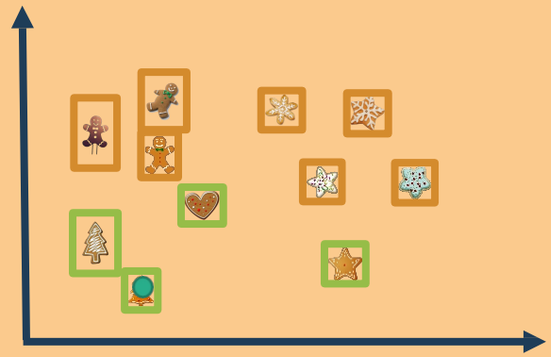
We specify the number K of clusters, in this case 2.

Steps 2 and 3 are repeated until the centers no longer move (stabilization).

2 disadvantages :

- we choose the number of classes K a priori.
- the partition depends on initialization

Step 1: Randomly designate 2 initial centers in the data, in this case 2 cookies.



Step 2: Each cookie is assigned to the nearest center. This gives 2 clusters, red and green.

K-means algorithm

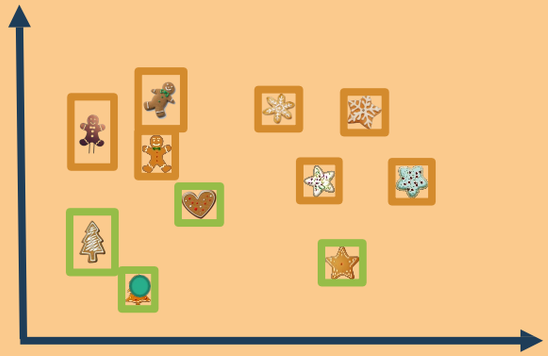
Fast clustering algorithm.

We specify the number K of clusters, in this case 2.

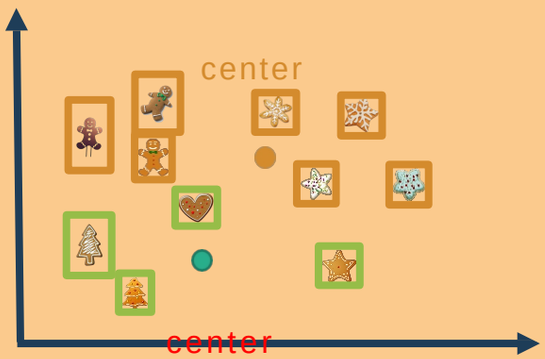
Steps 2 and 3 are repeated until the centers no longer move (stabilization).

2 disadvantages :
- we choose the number of classes K a priori.
- the partition depends on initialization

Step 1: Randomly designate 2 initial centers in the data, in this case 2 cookies.



Step 2: Each cookie is assigned to the nearest center. This gives 2 clusters, red and green.



Step 3: the centers are moved to the center (center of gravity) of their cluster

K-means algorithm

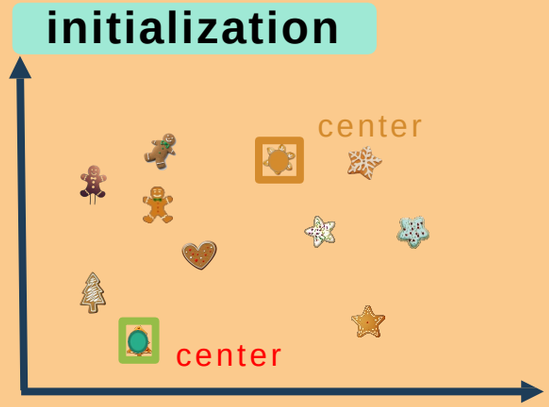
Fast clustering algorithm.

We specify the number K of clusters, in this case 2.

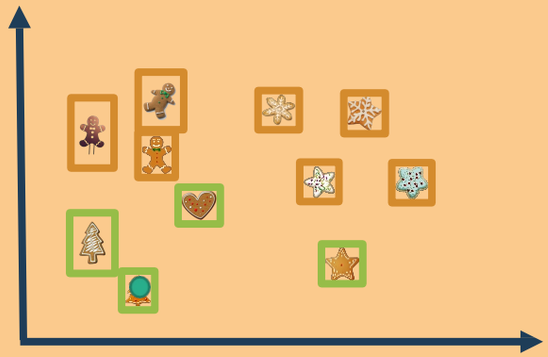
Steps 2 and 3 are repeated until the centers no longer move (stabilization).

2 disadvantages :

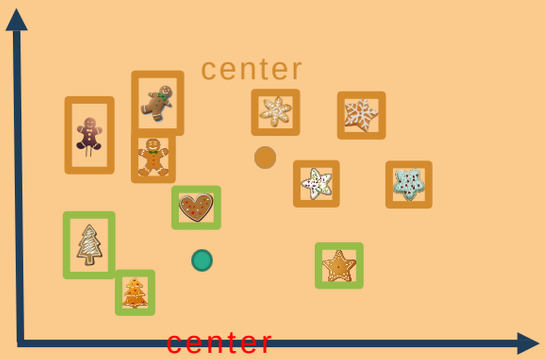
- we choose the number of classes K a priori.
- the partition depends on initialization



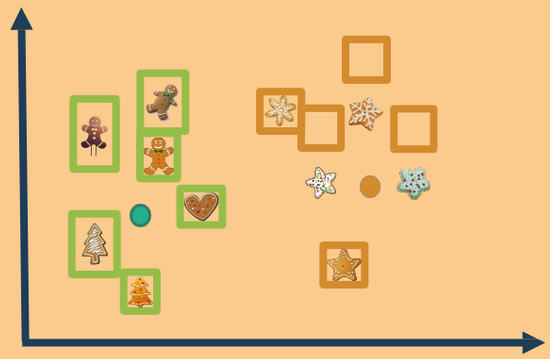
Step 1: Randomly designate 2 initial centers in the data, in this case 2 cookies.



Step 2: Each cookie is assigned to the nearest center. This gives 2 clusters, red and green.



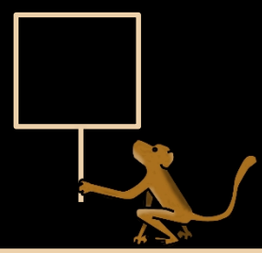
Step 3: the centers are moved to the center (center of gravity) of their cluster.

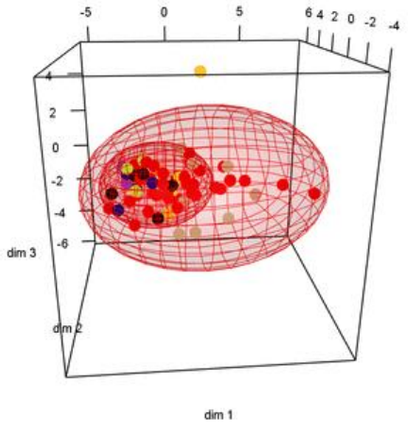


Stabilization:

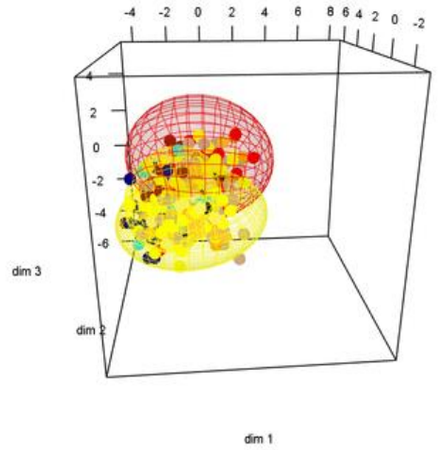
Final result

Using clusters to study monkey language

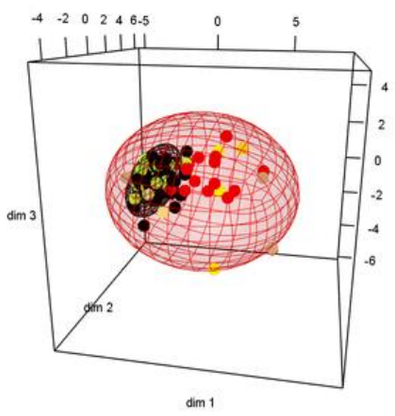




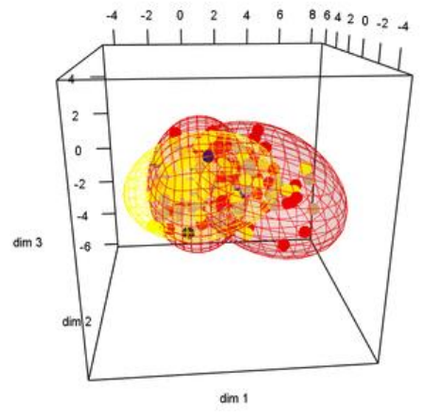
Japanese macaque



rhesus macaque



Tonkean macaque



crested macaque

Conclusions

- Visualization method
- Rigorous statistical method

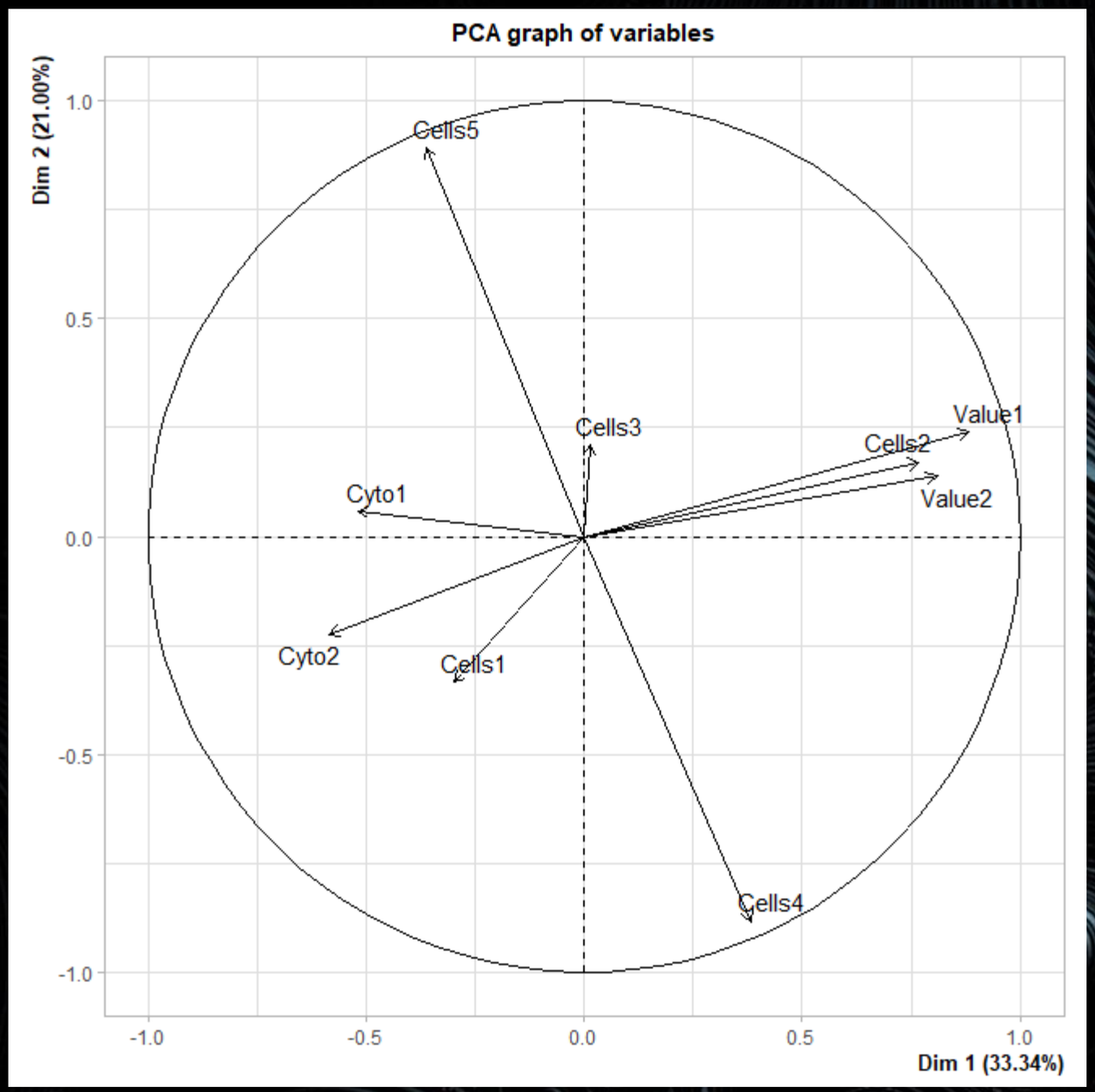
- Quick method

- R software
- FactoMineR package

In your field

Treatment	Cells1	Cells2	Cells3	Cells4	Cells5	Value1	Value2	Cyto1	Cyto2
NT	0.0235	165.00	4.07	74.0	5.84	1.0280	519.50000	0.1107	27.22
NT	0.0418	73.50	4.49	77.3	6.34	0.7905	438.54552	0.1368	33.58
NT	0.0340	55.50	3.51	90.5	1.18	0.5584	358.44663	0.0185	8.57
NT	0.0254	112.50	7.92	83.0	2.48	0.6869	465.80625	0.0158	6.76
NT	0.0257	92.50	5.96	80.0	5.62	0.5611	339.57500	0.0371	14.64

In your field



Take home message factor analysis

- Spatially close = similar
- Exploratory analysis