PCA Motivation

Predictive Modeling & Statistical Learning

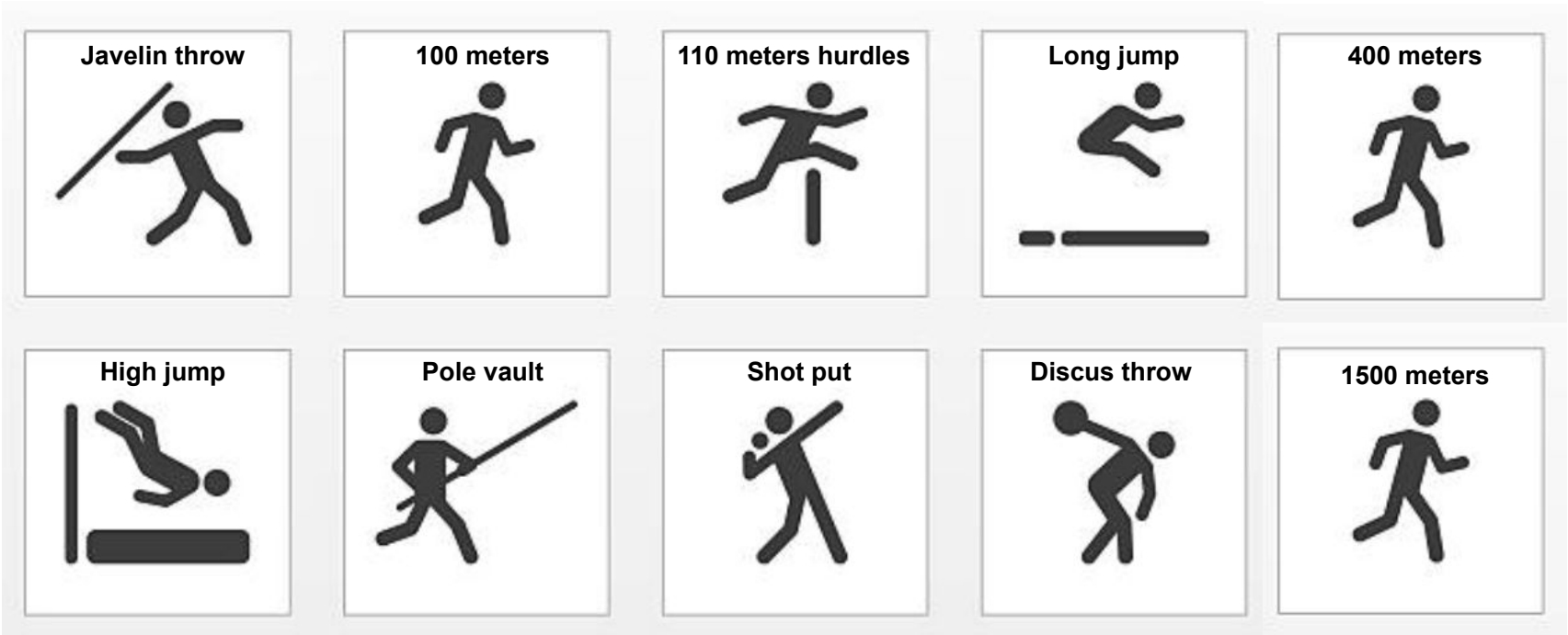
Gaston Sanchez

Introduction

Decathlon

- `decathlon` data
- from R package ‘‘`FactoMineR`’’
- 41 athletes, 13 variables

Decathlon Events



```
          100m  Long.jump  Shot.put  High.jump   400m
SEBRLE   11.04       7.58     14.83       2.07  49.81
CLAY     10.76       7.40     14.26       1.86  49.37
KARPOV   11.02       7.30     14.77       2.04  48.37
BERNARD  11.02       7.23     14.25       1.92  48.93
YURKOV   11.34       7.09     15.19       2.10  50.42
```

```
         110m.hurdle  Discus  Pole.vault  Javeline  1500m
SEBRLE         14.69   43.75        5.02     63.19  291.7
CLAY           14.05   50.72        4.92     60.15  301.5
KARPOV         14.09   48.95        4.92     50.31  300.2
BERNARD        14.99   40.87        5.32     62.77  280.1
YURKOV         15.31   46.26        4.72     63.44  276.4
```

Exploratory Data Analysis

Let's focus on the following variables:

- `100m`
- `Long.jump`
- `Shot.put`
- `High.jump`
- `400m`
- `110m.hurdle`
- `Discus`
- `Pole.vault`
- `Javeline`
- `1500m`

EDA: Objects and Variables Perspectives

Data Perspectives

We are interested in analyzing a data set from both perspectives: **objects** and **variables**

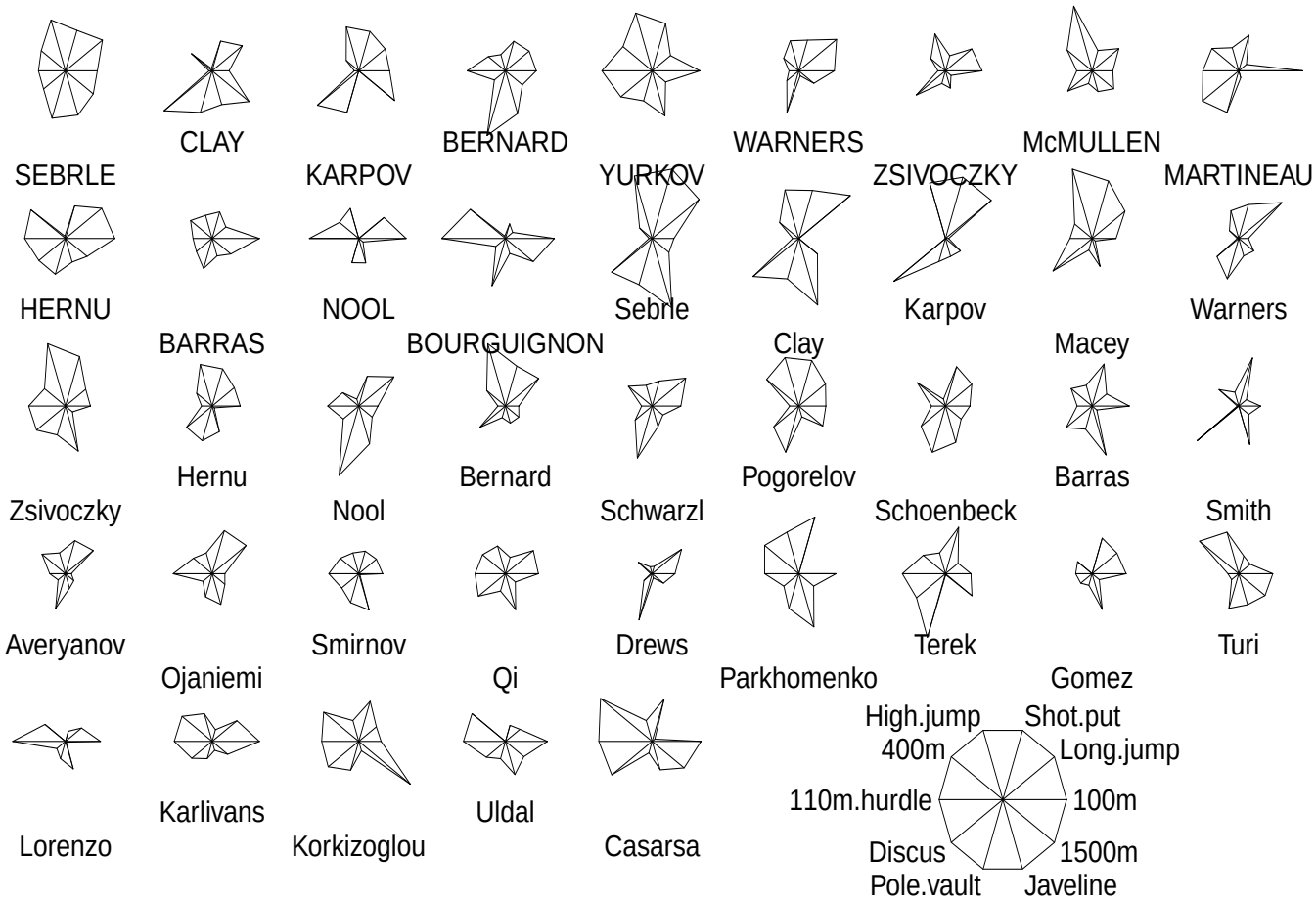
At its simplest we are interested in 2 fundamental purposes:

- Study resemblance among individuals
  (resemblance among athletes)
- Study relationship among variables
  (relationship among events statistics)

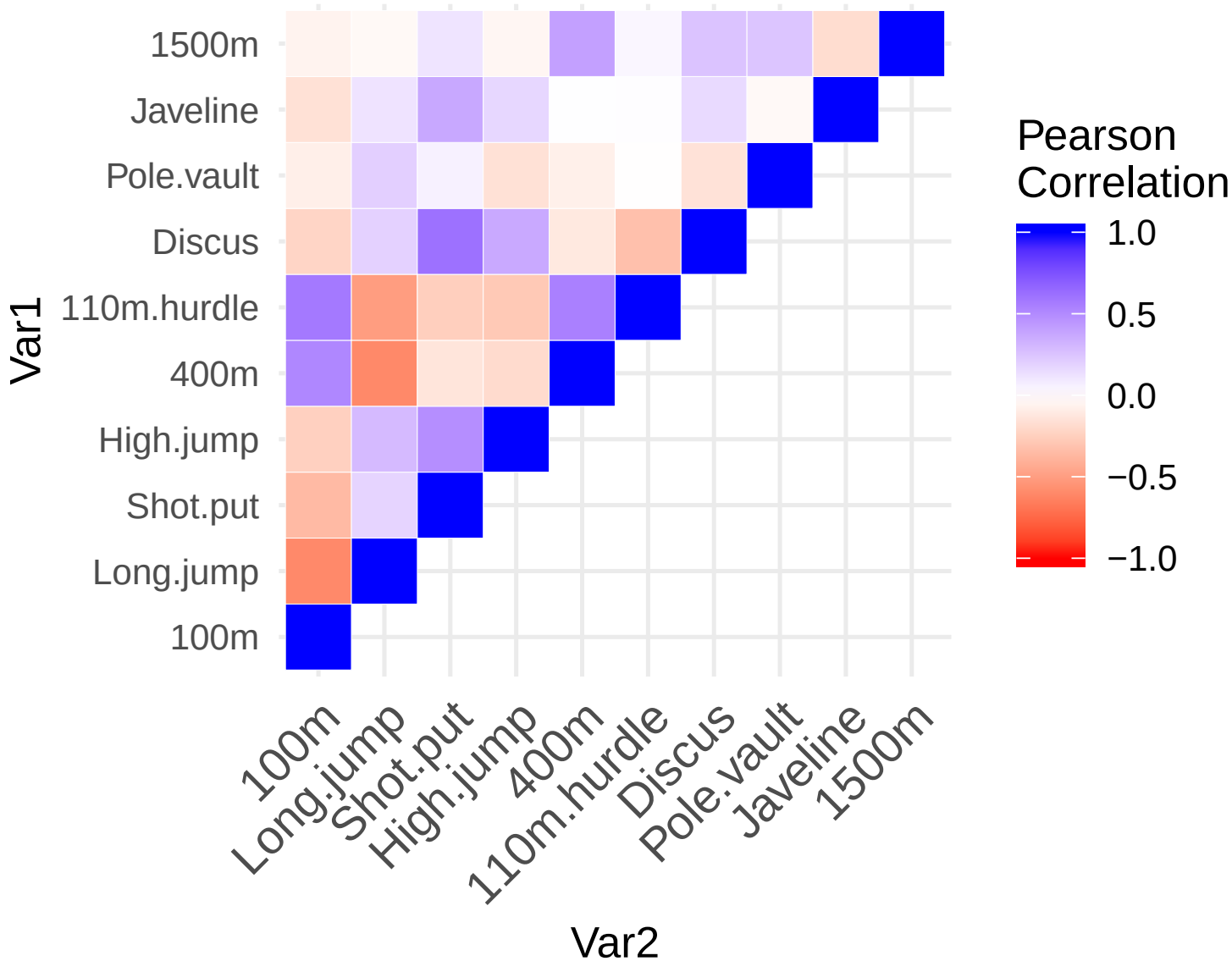
Exploration

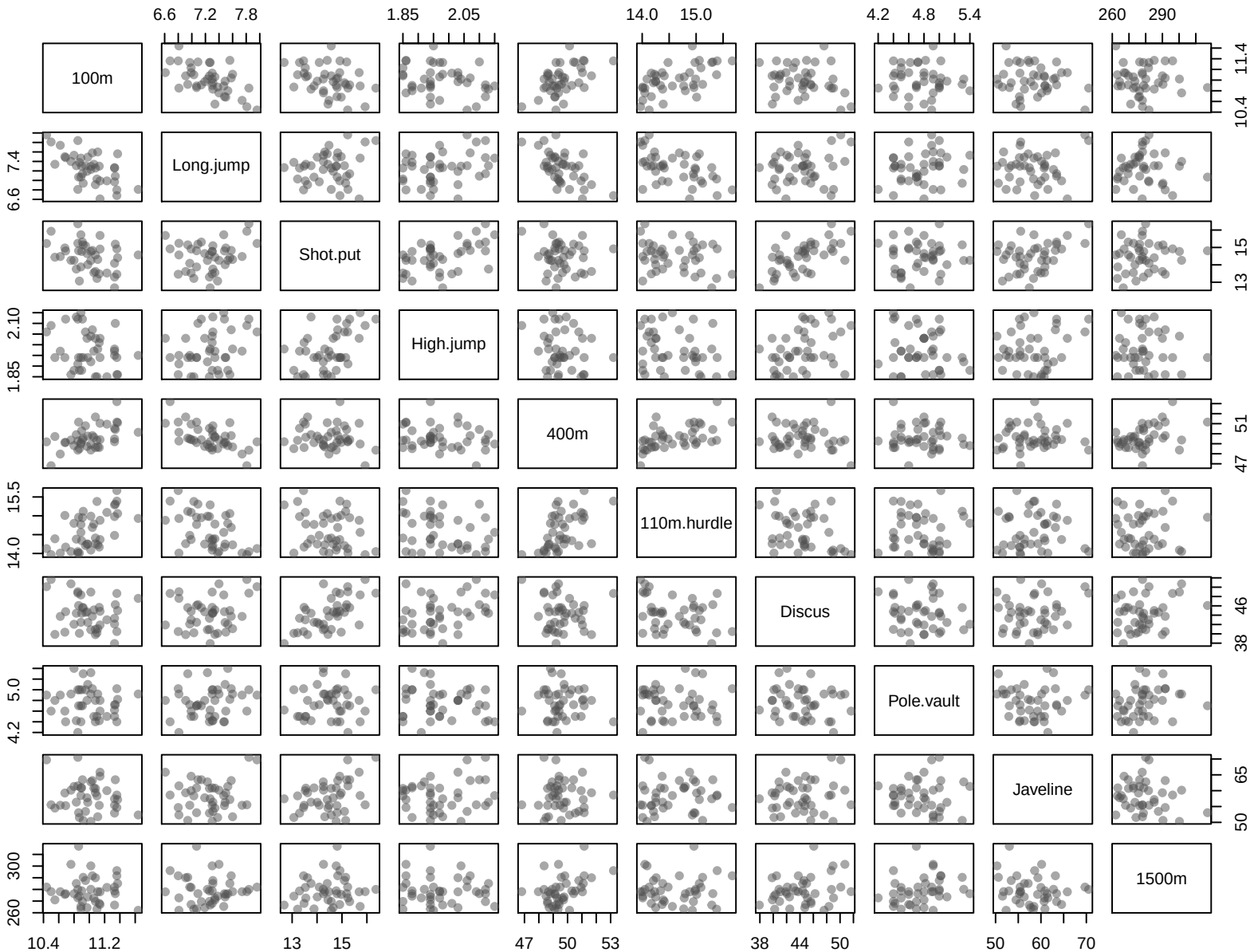
Likewise, we can explore variables at different stages:

- Univariate: one variable at a time
- Bivariate: two variables simultaneously
- Multivariate: multiple variables

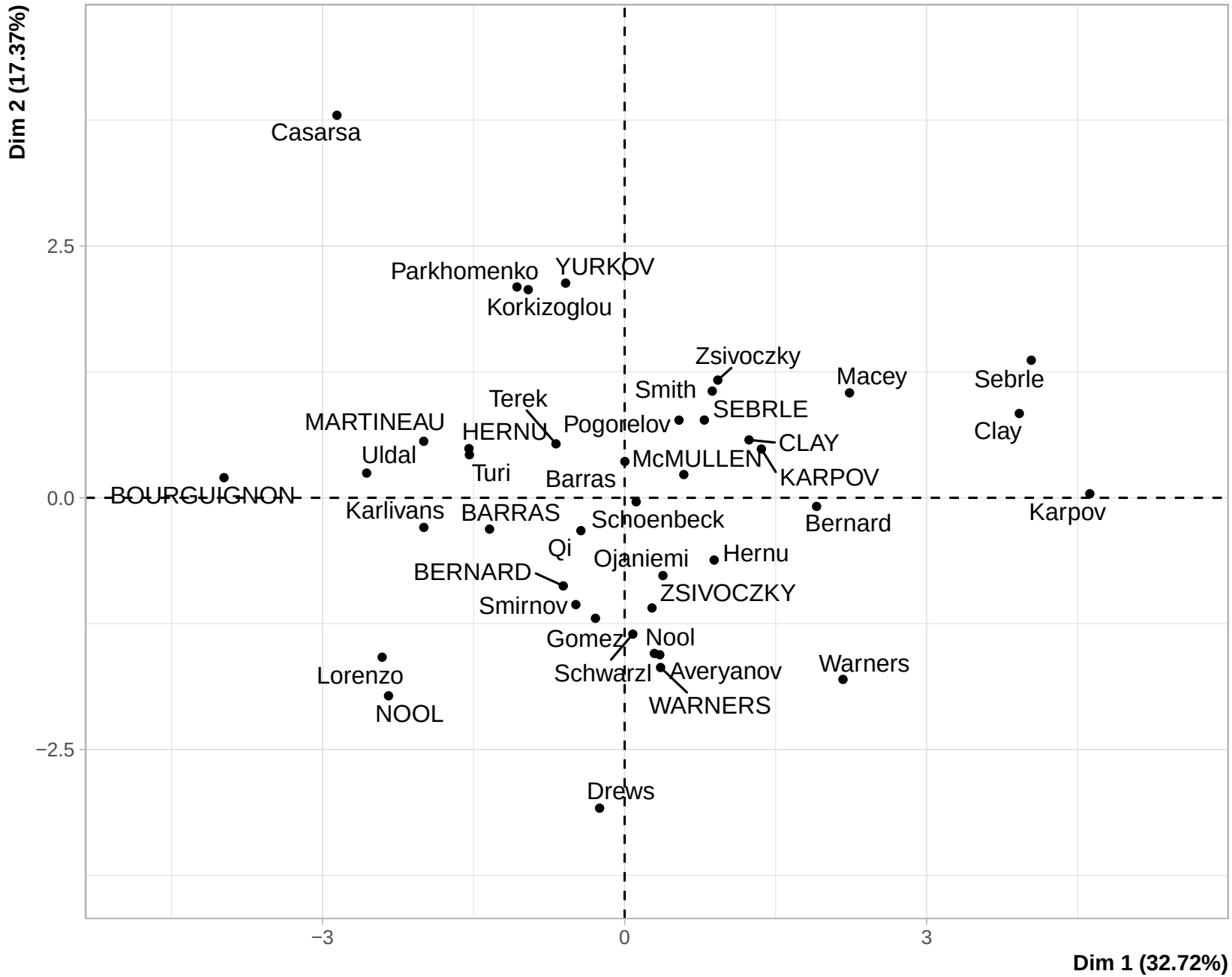

SEBRLE
CLAY
KARPOV
BERNARD
YURKOV
WARNERS
ZSIVOCZKY
McMULLEN
MARTINEAU
HERNU
BARRAS
NOOL
BOURGUIGNON
Sebrle
Clay
Karpov
Macey
Warners
Zsivoczky
Hernu
Nool
Bernard
Schwarzl
Pogorelov
Schoenbeck
Barras
Smith
Averyanov
Ojaniemi
Smirnov
Qi
Drews
Parkhomenko
Terek
Gomez
Turi
Lorenzo
Karlivans
Korkizoglou
Uldal
Casarsa
High.jump
Shot.put
400m
Long.jump
110m.hurdle
100m
Discus
1500m
Pole.vault
Javeline

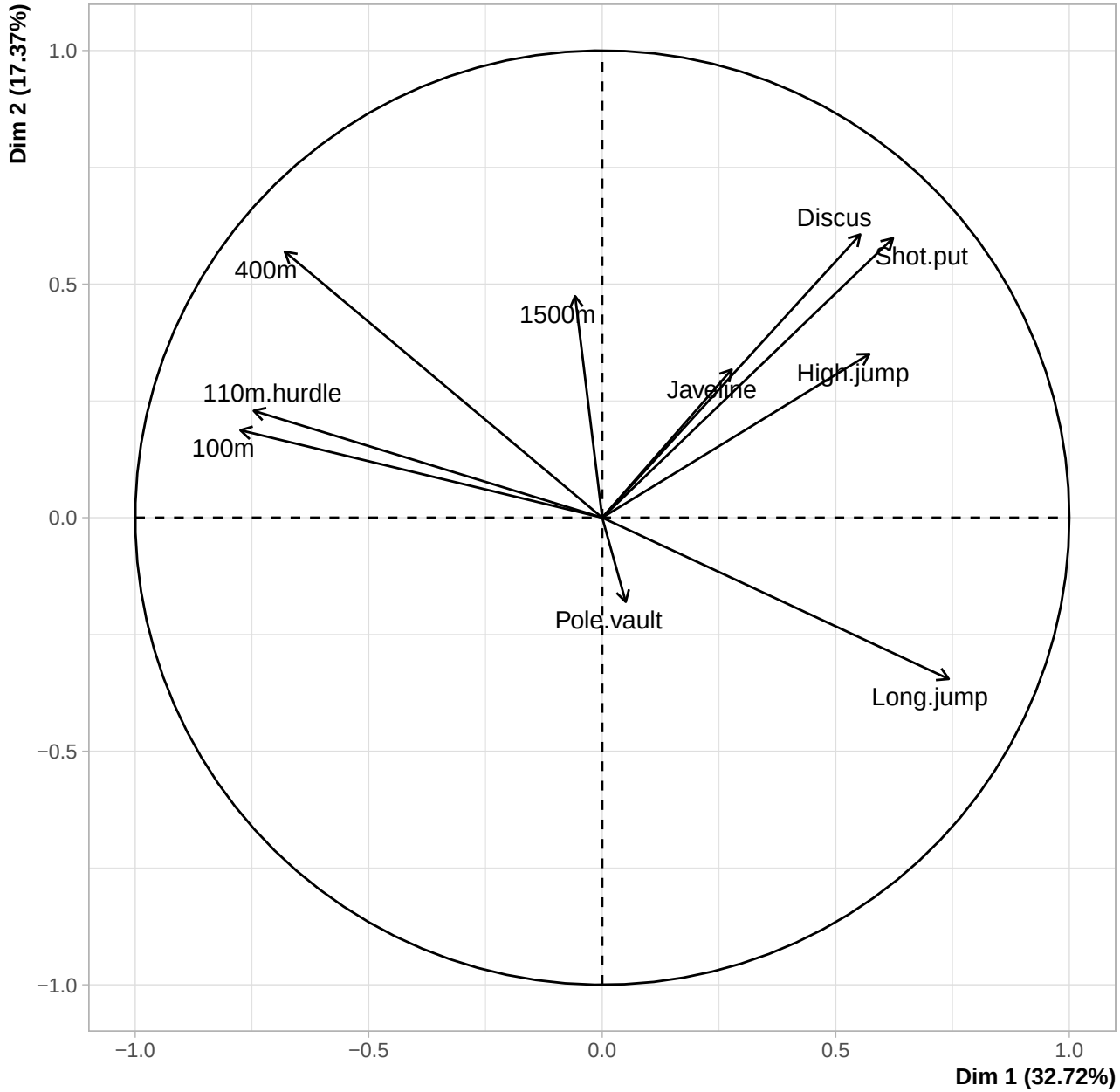
Correlation heatmap

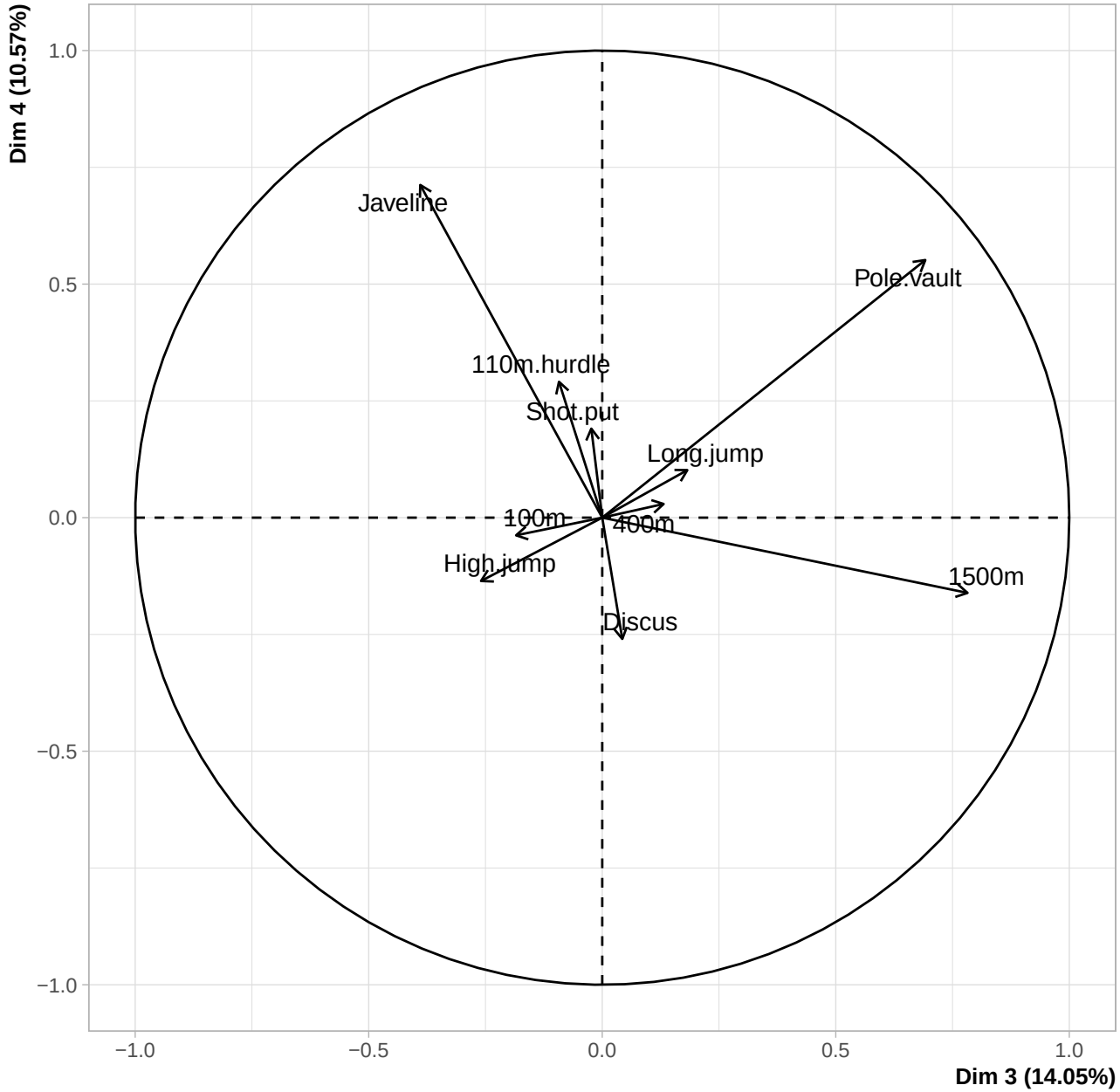



100m
Long.jump
Shot.put
High.jump
400m
110m.hurdle
Discus
Pole.vault
Javeline
1500m

What if we could get a better low-dimensional summary of the data?


Dim 2 (17.37%)
2.5
0.0
−2.5
−3
0
3
Dim 1 (32.72%)
Casarsa
Parkhomenko
YURKOV
Korkizoglou
Zsivoczky
Macey
Sebrle
Smith
Terek
SEBRLE
MARTINEAU
HERNU
Pogorelov
Clay
CLAY
Uldal
McMULLEN
Turi
KARPOV
Barras
BOURGUIGNON
Karlivans
BARRAS
Schoenbeck
Bernard
Karpov
Qi
Ojaniemi
Hernu
BERNARD
ZSIVOCZKY
Smirnov
Gomez
Nool
Warners
Lorenzo
Schwarzl
Averyanov
WARNERS
NOOL
Drews


Dim 2 (17.37%)
1.0
0.5
0.0
−0.5
−1.0
Discus
Shot.put
400m
1500m
High.jump
Javeline
110m.hurdle
100m
Pole.vault
Long.jump
−1.0
−0.5
0.0
0.5
1.0
Dim 1 (32.72%)


Dim 4 (10.57%)
1.0
0.5
0.0
−0.5
−1.0
Javeline
Pole.vault
110m.hurdle
Shot.put
Long.jump
100m
400m
High.jump
1500m
Discus
−1.0
−0.5
0.0
0.5
1.0
Dim 3 (14.05%)