# Principal Component Analysis

## A Story with Wines
## Guest Star: The Eigen Fairy

Hervé Abdi & Ju-Chi Yu

# Outline

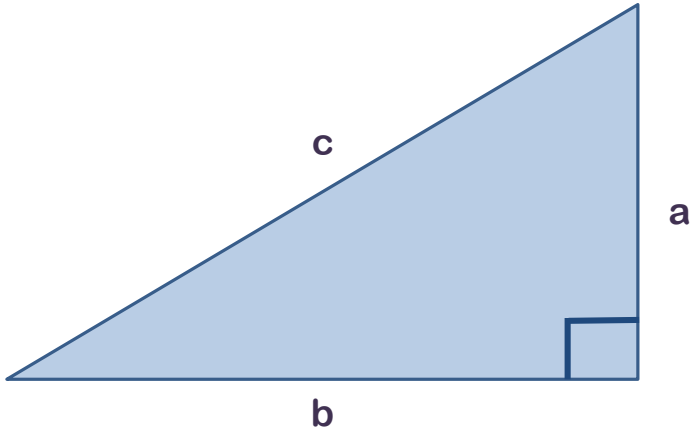
## Today's

- The core of it all:
- Principal Component Analysis (PCA)

# What do you really need to know

- The **THEOREM:**
- **Yes, this one!**
- **The Pythagorean Theorem**

$$a^2 + b^2 = c^2$$

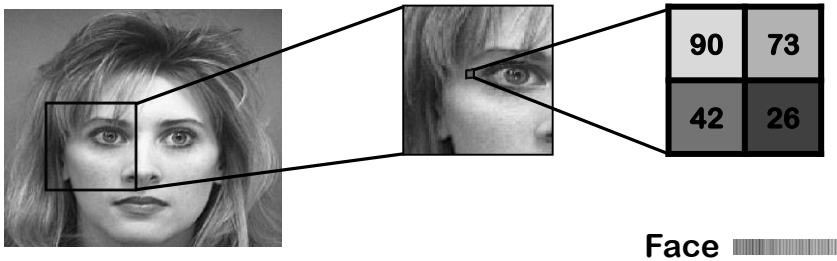


# What are Multivariate Data?

- Surveys

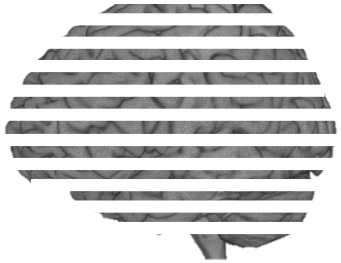
| Personality | 5 | 9 | ... | 0 | 3 | ... | 10 | 8 | ... | 5 |
|---|---|---|---|---|---|---|---|---|---|---|

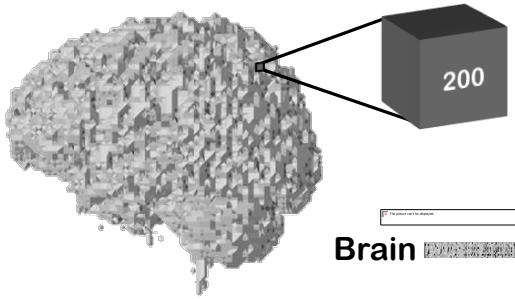
- Images

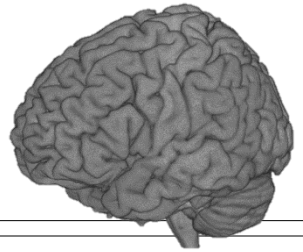



Face

- Brains





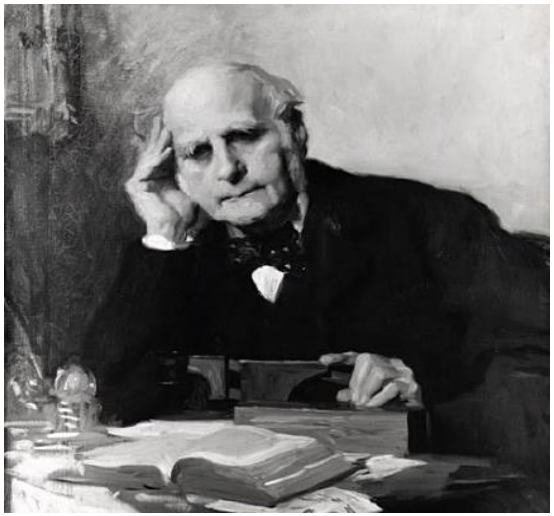
Brain

# Principal Component Analysis

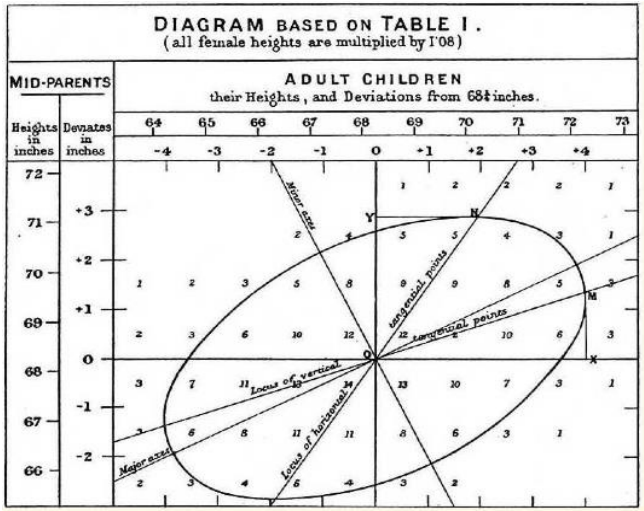
- You already did it today!
- Huh?
- Example 1.
- Example 1. Google (original page rank)
- Example 2
- Example 2. Music (.mp3)
- Example 3
- Example 3. Netflix
- Example 4, 5, ….

# Principal Component Analysis

- **Oldest multivariate technique**
  - Idea: Cauchy (1829); Galton (1859); Pearson (1902)
  - AKA: Eigen-analysis; singular value decomposition, Hotelling transform, Karhunen-Loève decompositions, …



Sir Francis Galton
(1822-1911)




(A Kind of) PCA Comparing Mid-Parental to Adult Children Height

# Principal Component Analysis

- Oldest multivariate technique
  - Idea: Cauchy (1829); Galton (1859); Pearson (1902)
  - AKA: Eigen-analysis; singular value decomposition, Hotelling transform, Karhunen-Loève decompositions, …
- But really do-able only with computers
- Goals of PCA:
  - Extracts important information from a data table
  - Represents the information as nicer variables
  - Replaces “nasty numbers” by “nice maps”
    - For both observations and variables

# Working with Multivariate Data

- **Common characteristics**
  - Often more variables than observations ($N << P$)
    - Traditional analyses cannot handle this
  - Inference: usual assumptions may not hold
- **Goal of multivariate analysis techniques**
  - Extract most important information
  - Present it well

# PCA Example with 20 wines and 2 variables

# 20 Wines

## Meet the Expert

Our Expert tasted 20 (+1) wines

Rated them on 2 scales (0 to 20):

*Sugar (Sweet)*

*Astringency*

10 Wines were from the US (Zinfandel)

10 Wines were from France (Cabernet + Merlot)

Later on, rated the wines on *Bitter* and *Acidic*

Commented:

Some wines were *Fruity*, some were *Woody*

## Hypothesis:

Sugar & Astringent inversely perceived

# 20 Wines

| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Lisrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Hordow | 8. St. Estèphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emillion | 20. Tomasello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Lisrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Hordow | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14. Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| F: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| W: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |

# 20 Wines: All data

| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Listrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Hordow | 8. St. Estèphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emillion | 20. Tomasello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Listrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Hordow | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14. Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| Y: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| W: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |
| Fruity | | | F | | F | | | | | F | F | | | F | F | F | | | | |
| Woody | | | | | | | | | | | | W | W | | | | | W | W | W |
| Acid | 2 | 5 | 1 | 1 | 9 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 9 | 14 | 2 | 1 | 2 | 1 | 1 |
| Bitter | 8 | 3 | 16 | 3 | 11 | 1 | 1 | 9 | 1 | 8 | 3 | 2 | 9 | 12 | 10 | 2 | 1 | 10 | 7 | 3 |

# 20 Wines

| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Lisrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Hordow | 8. St. Estèphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emillion | 20. Tomasello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Lisrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Hordow | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14. Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| F: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| W: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |

## Hypothesis:

## Sugar & Astringent inversely perceived

# WE ARE LOST WITH NUMBERS!

| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Listrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Hordew | 8. St. Estèphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emilion | 20. Tommaello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Listrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Hordew | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14 Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| F: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| W: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |

Numbers

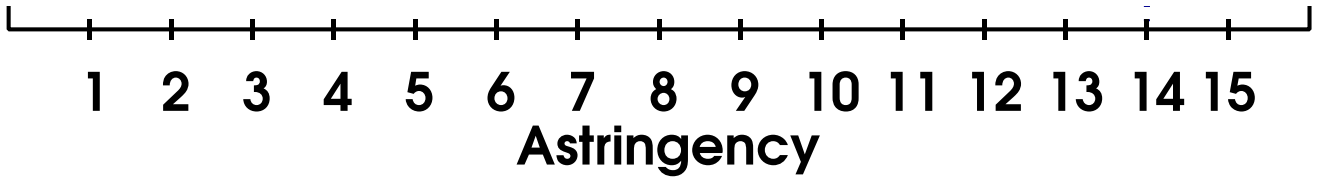
# So: Make a Picture!

| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Listrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Horlow | 8. St Estéphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emillion | 20. Tomasello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Listrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Horlow | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14. Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| *Y*: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| *W*: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |

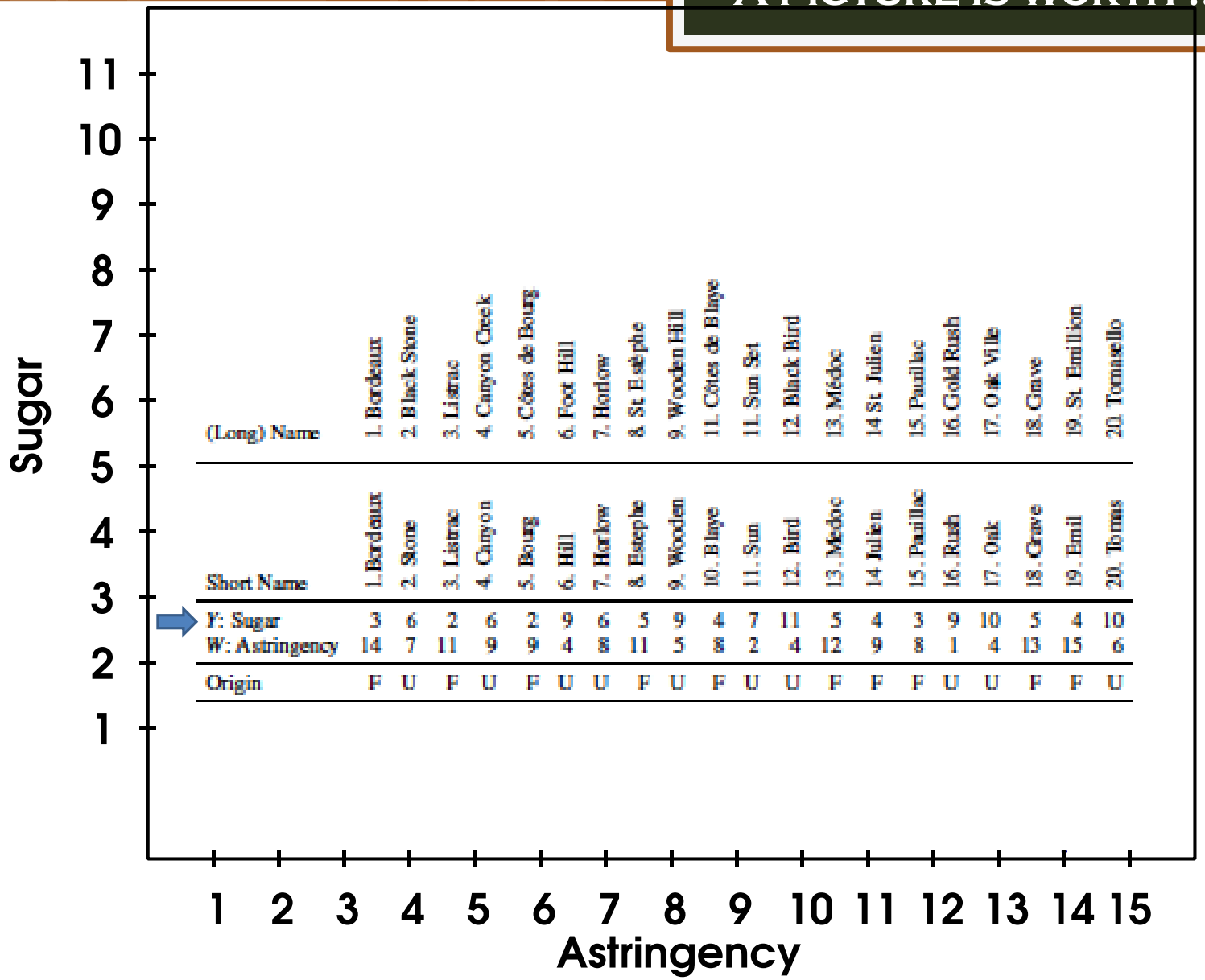
| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Listrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Horlow | 8. St Estéphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emillion | 20. Tomasello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Listrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Horlow | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14. Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| *Y*: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| *W*: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |

# A PICTURE IS WORTH …

| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Listrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Horlow | 8. St Estéphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emillion | 20. Tomasello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Listrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Horlow | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14. Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| Y: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| W: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |

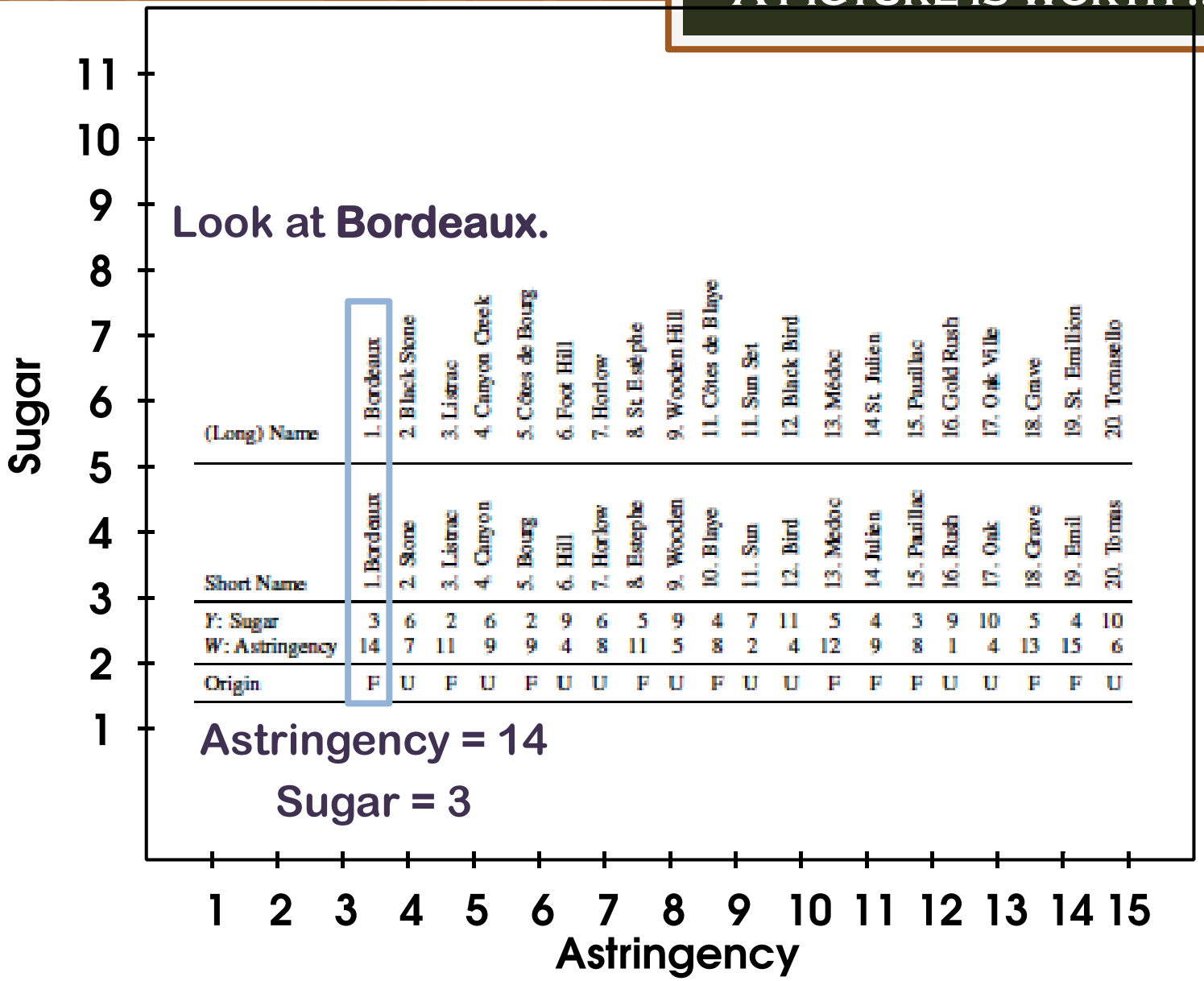


# A picture is worth …



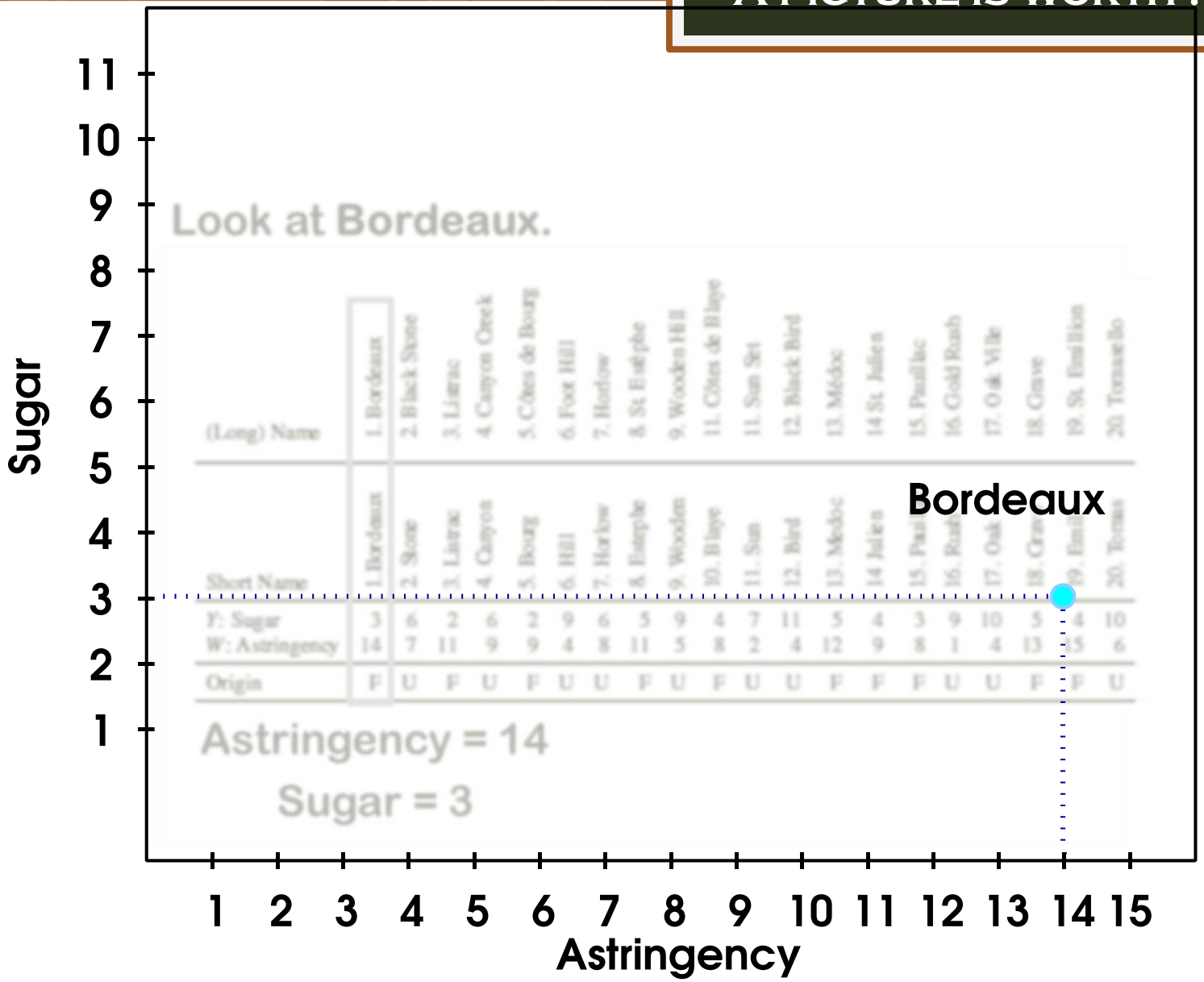

| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Listrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Horlow | 8. St Estéphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emillion | 20. Tomasello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Listrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Horlow | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14. Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| Y: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| W: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |

# A picture is worth …

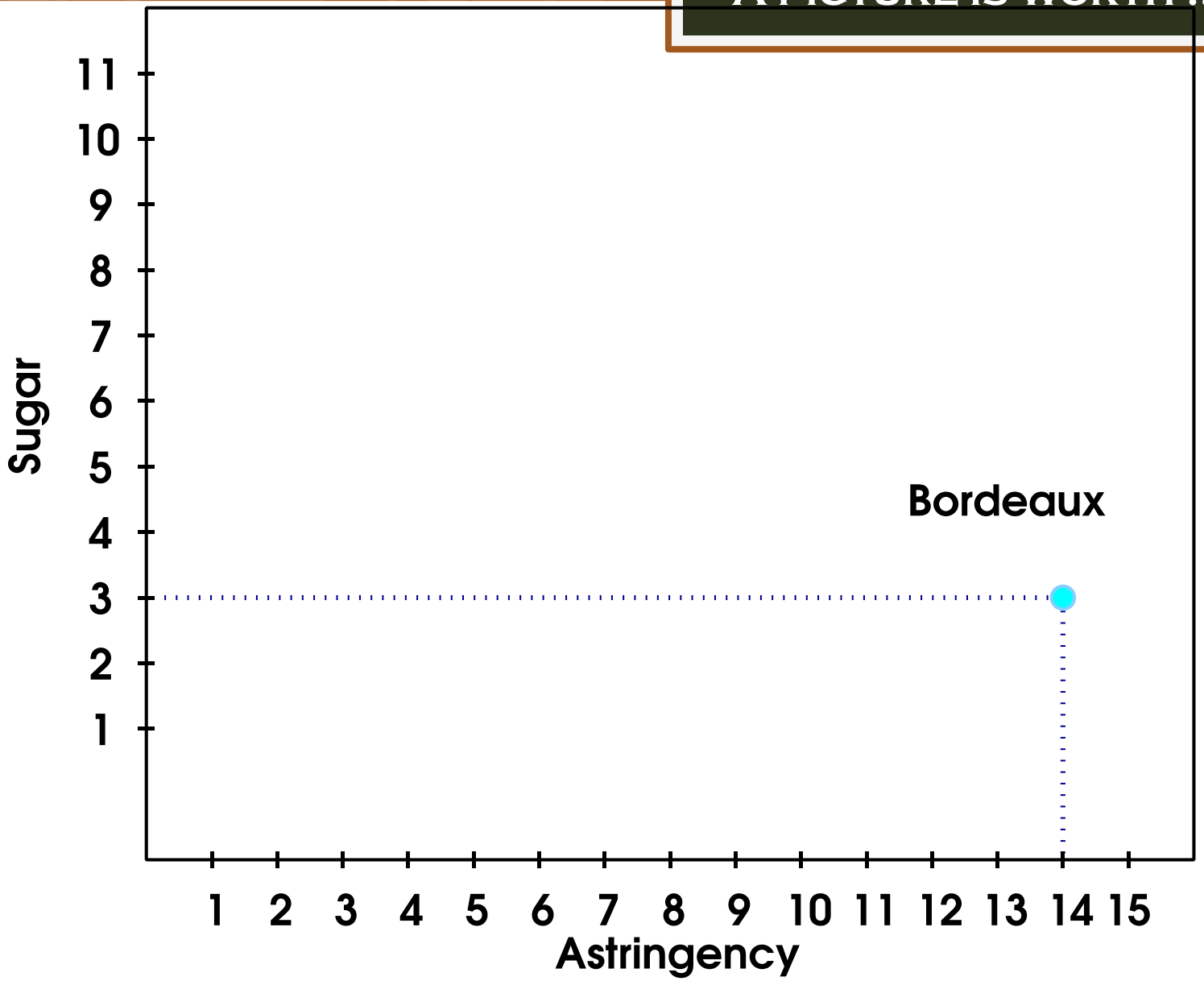



| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Listrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Horlow | 8. St Estéphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emillion | 20. Tomasello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Listrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Horlow | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14. Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| Y: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| W: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |

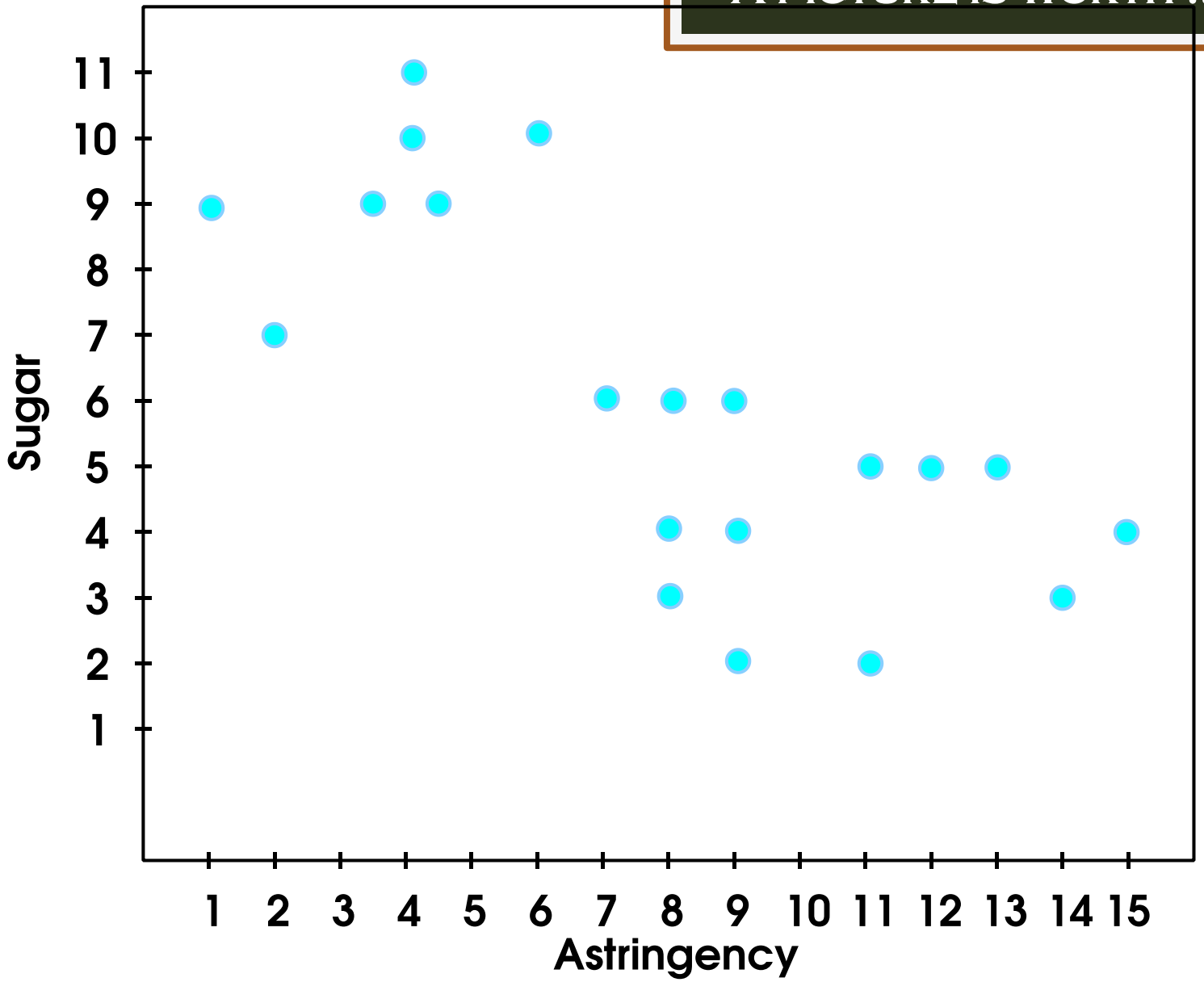
# A picture is worth …



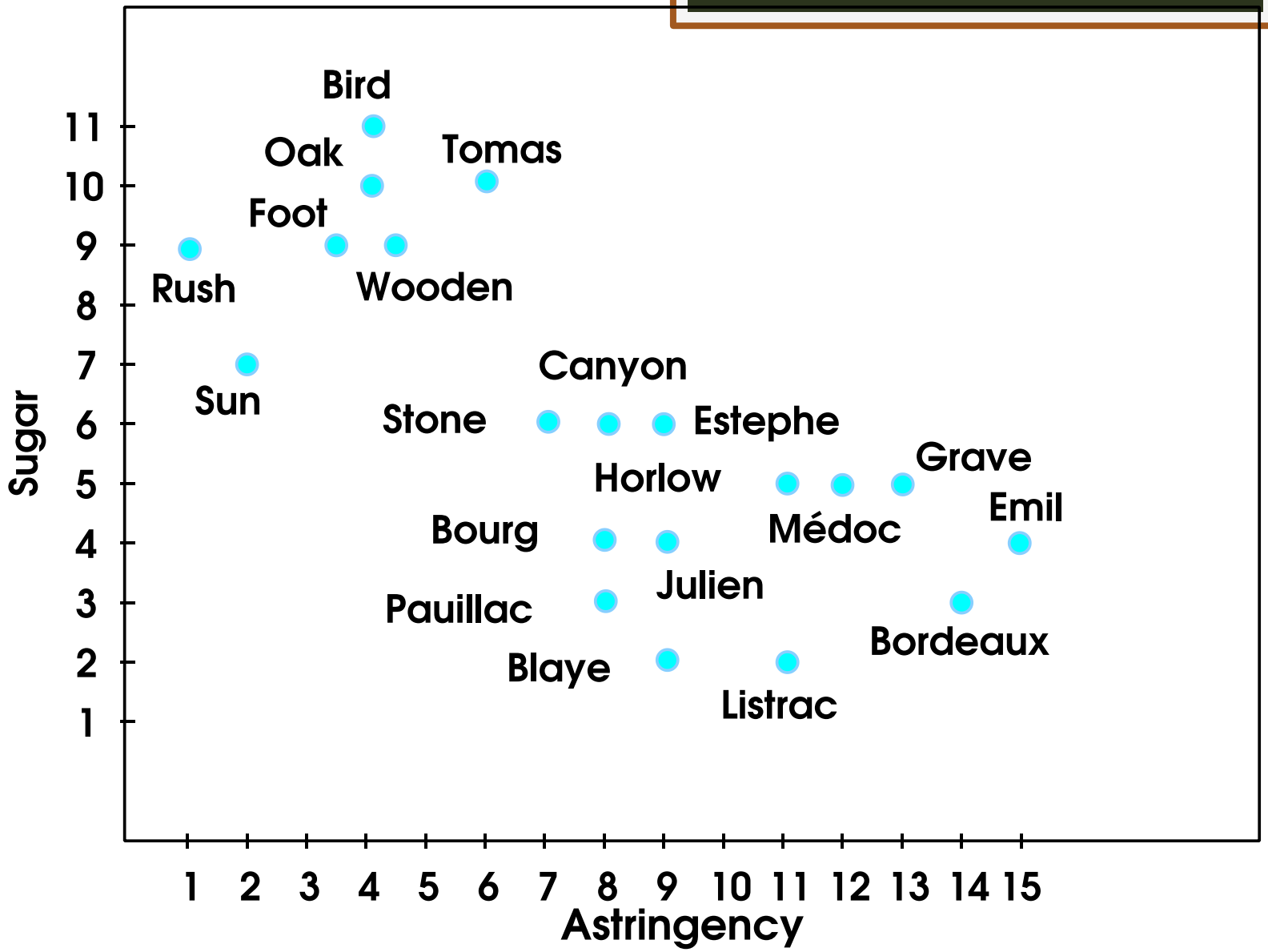
# A PICTURE IS WORTH ...



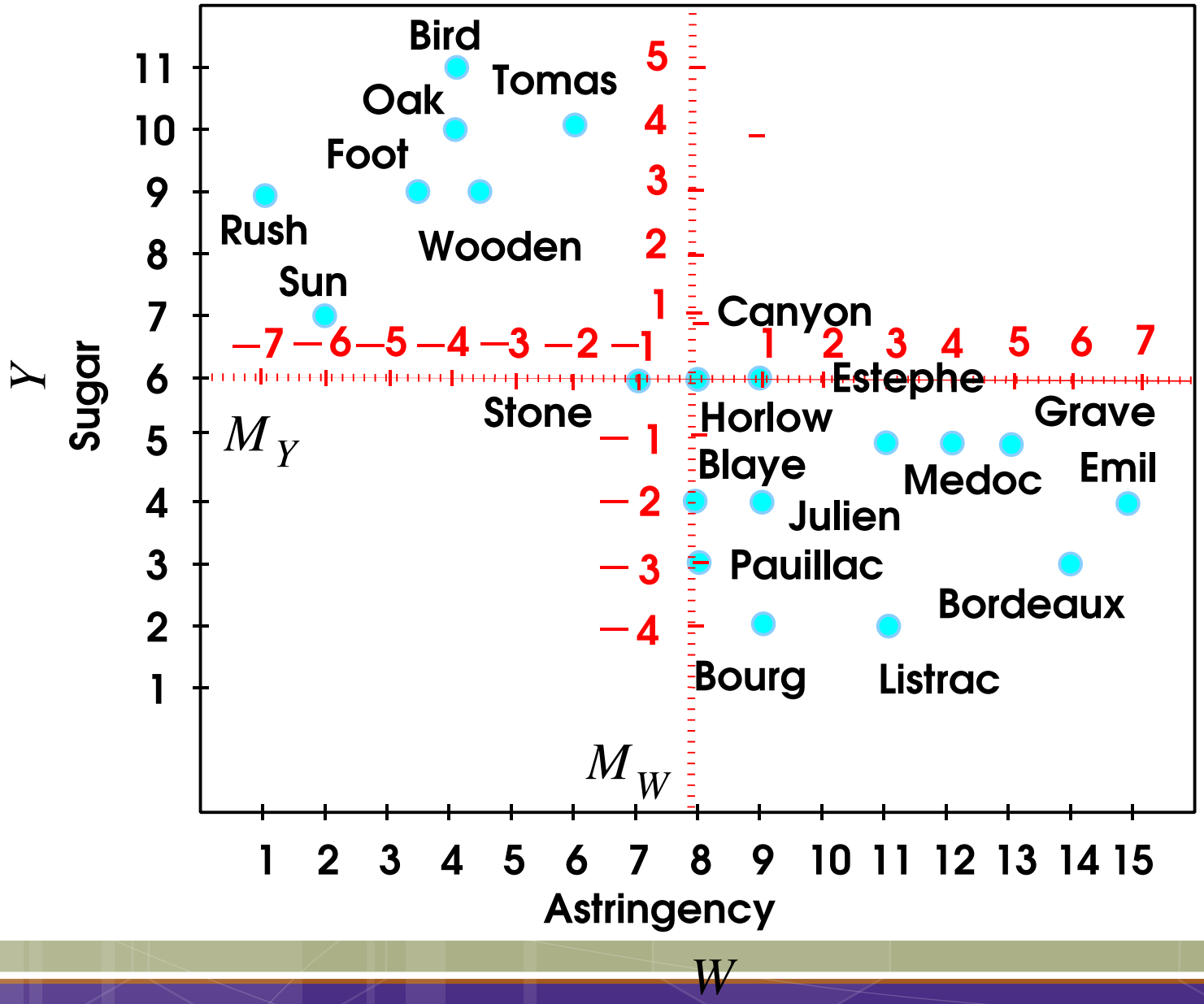
# A PICTURE IS WORTH ...



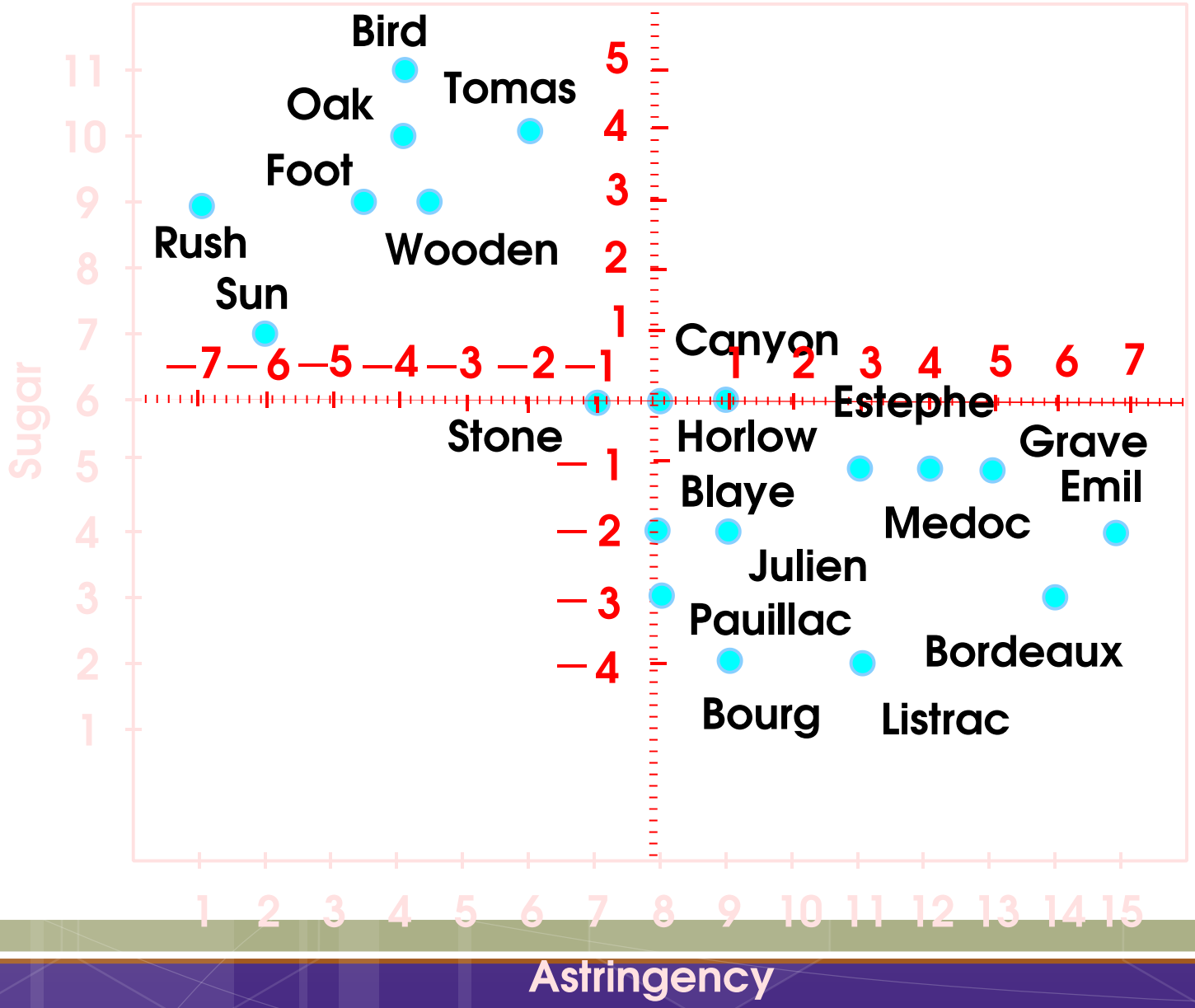
# A PICTURE ...


Bird
Oak
Tomas
Foot
Rush
Wooden
Sun
Canyon
Stone
Estephe
Grave
Horlow
Emil
Bourg
Médoc
Julien
Pauillac
Bordeaux
Blaye
Listrac
Sugar
11
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
Astringency

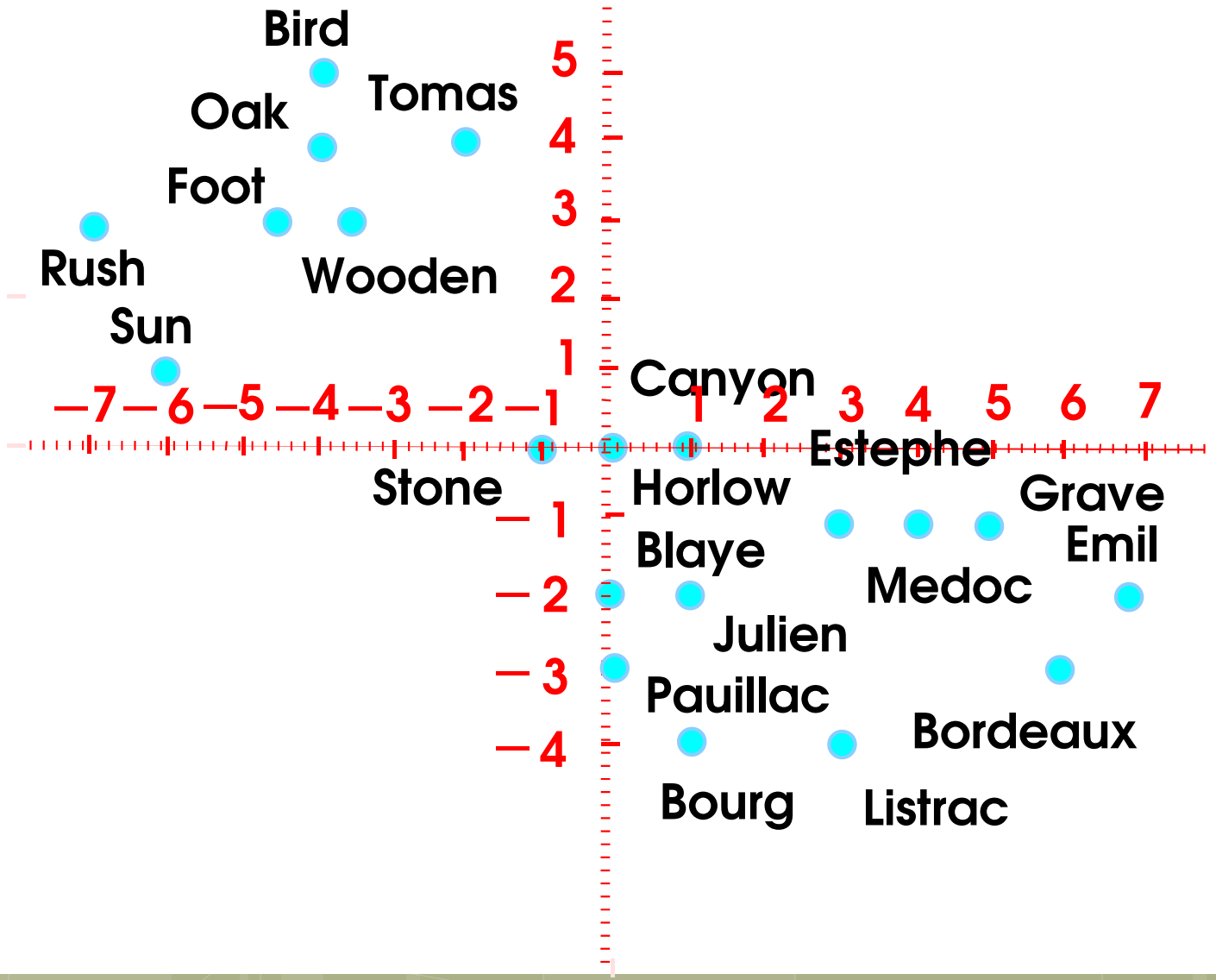
# A PICTURE ...


Bird
Tomas
Oak
Foot
Rush
Wooden
Sun
Canyon
Estephe
Stone
Horlow
Grave
Blaye
Medoc
Emil
Julien
Pauillac
Bordeaux
Bourg
Listrac
$M_Y$
$M_W$
Y
Sugar
Astringency
W

# A PICTURE ...


Bird
Tomas
Oak
Foot
Rush
Wooden
Sun
Canyon
Stone
Horlow
Estephe
Grave
Blaye
Emil
Medoc
Julien
Pauillac
Bordeaux
Bourg
Listrac
Sugar
Astringency

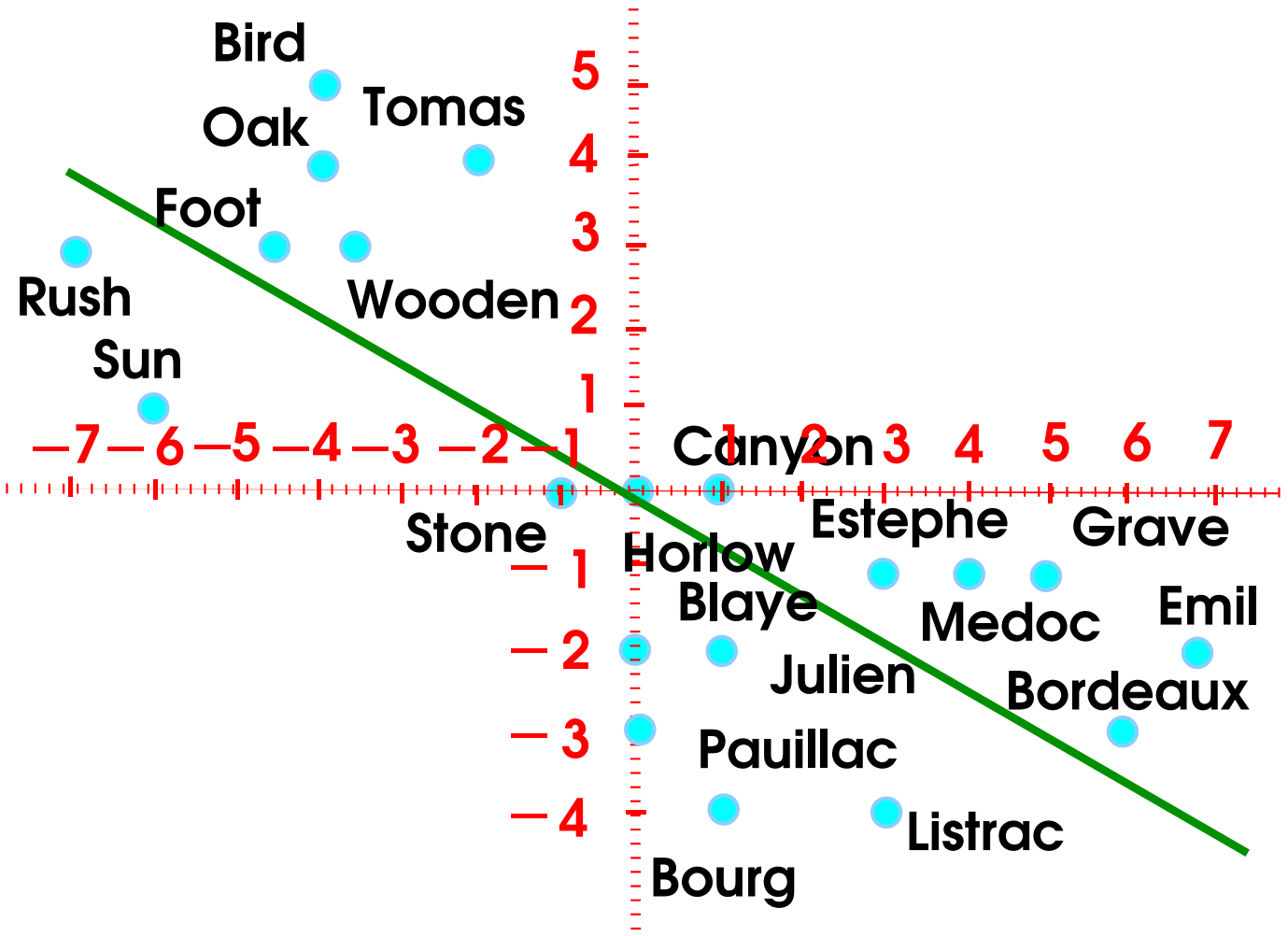
# A picture ... Centering



# A picture ... Centering

Could we get a line passing in the middle of this cloud?

# A PICTURE ... PC1



# A picture ... Centering

How to do it?

# Eyeballing!

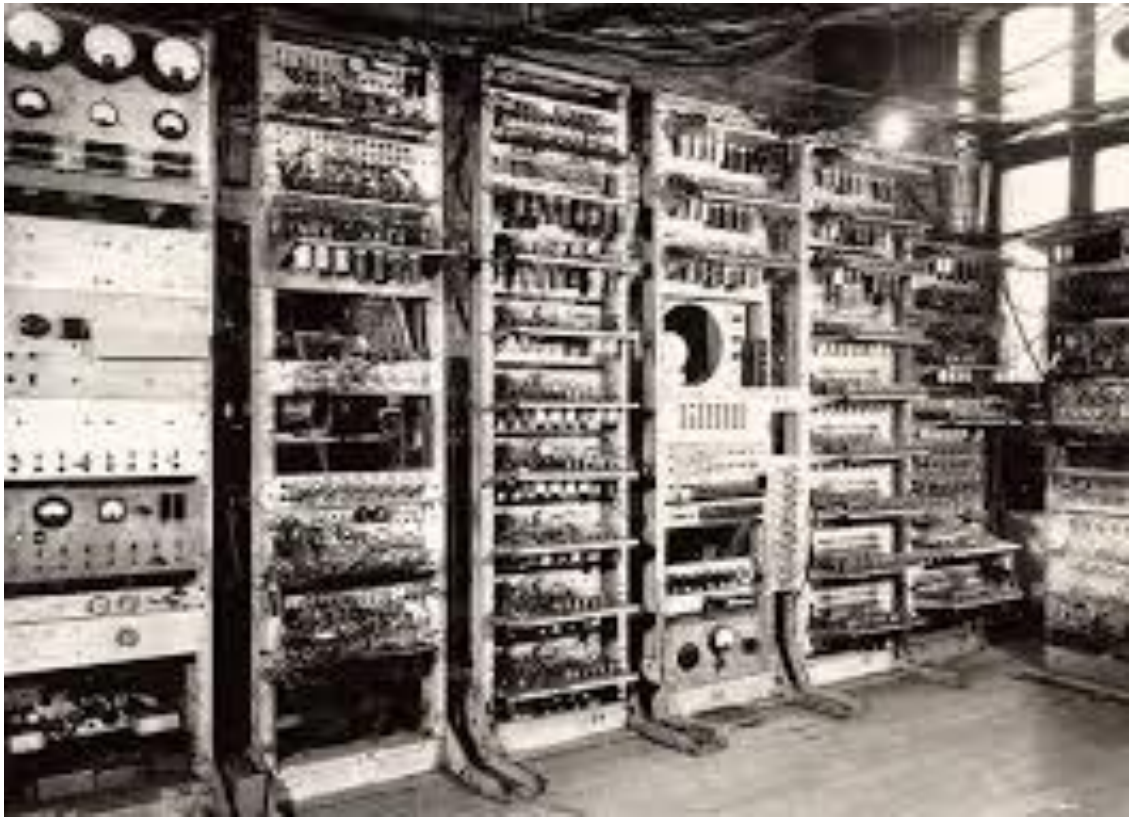
## How to do it?

# For the handy ones …

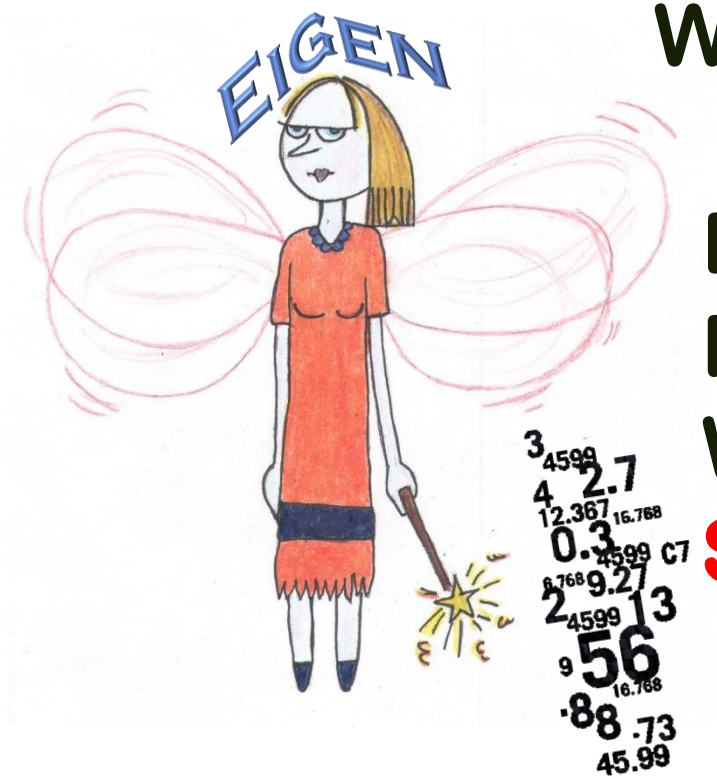
## How to do it?

## The old way …

# Some Computing ...



# Or Even Math Magic! The Eigen Fairy




## What to ask for?

Eigen Fairy, Please Find the rotation With **smallest sum of squared distances**

$$D^2 = A^2 + B^2$$

# PYTHAGOREAN MAGIC:


## SMALLEST SUM OF SQUARED DISTANCES

## SAME AS

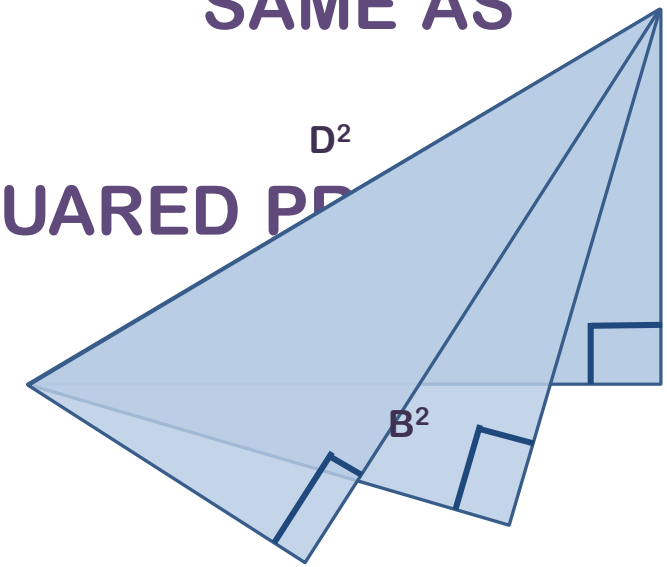
## LARGEST SQUARED PROJECTIONS (VARIANCE)

$D^2 = A^2 + B^2$

# PYTHAGOREAN MAGIC:
## SMALLEST SUM OF SQUARED DISTANCES

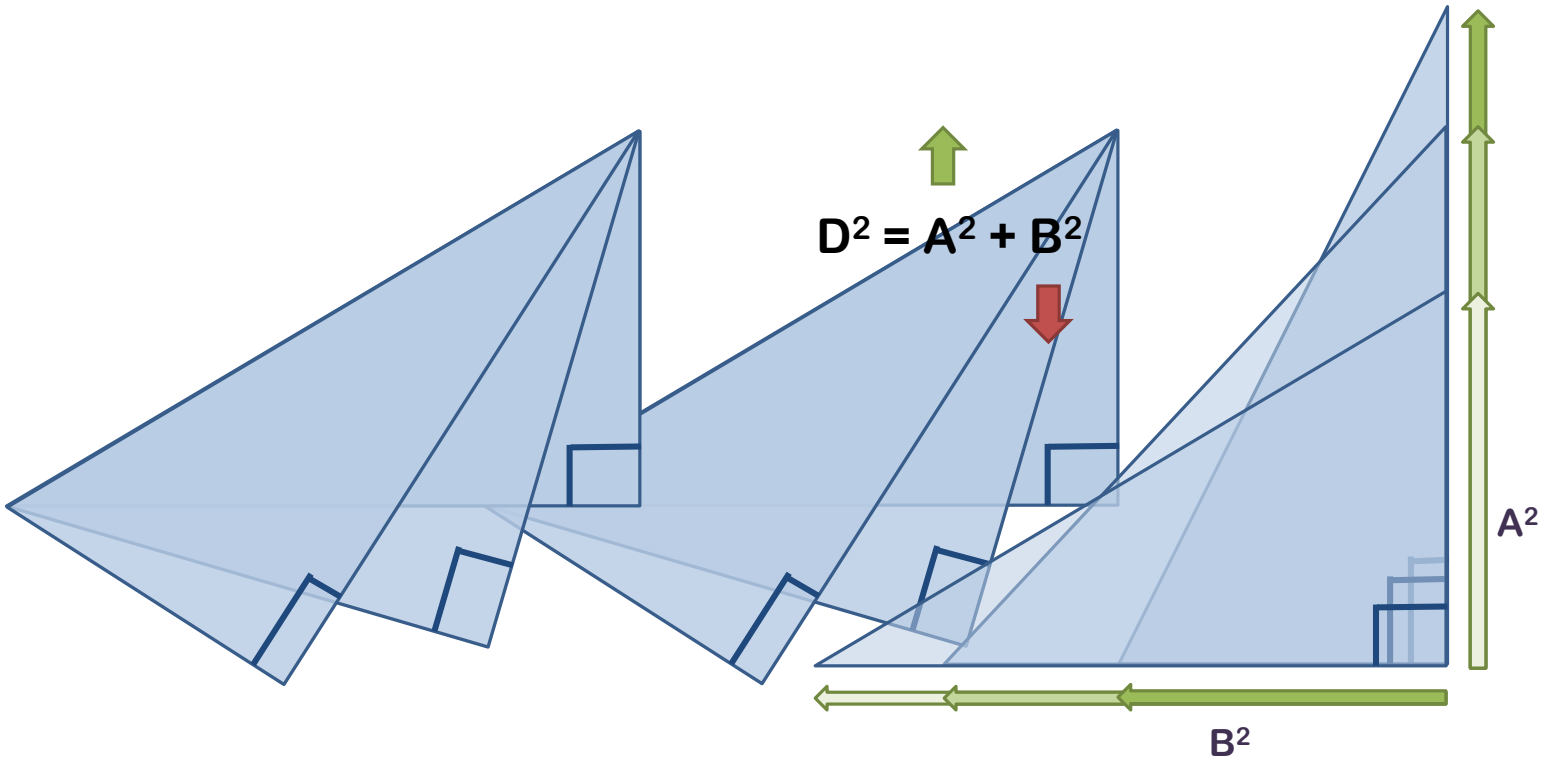
SAME AS

LARGEST SQUARED PR NS (VARIANCE)

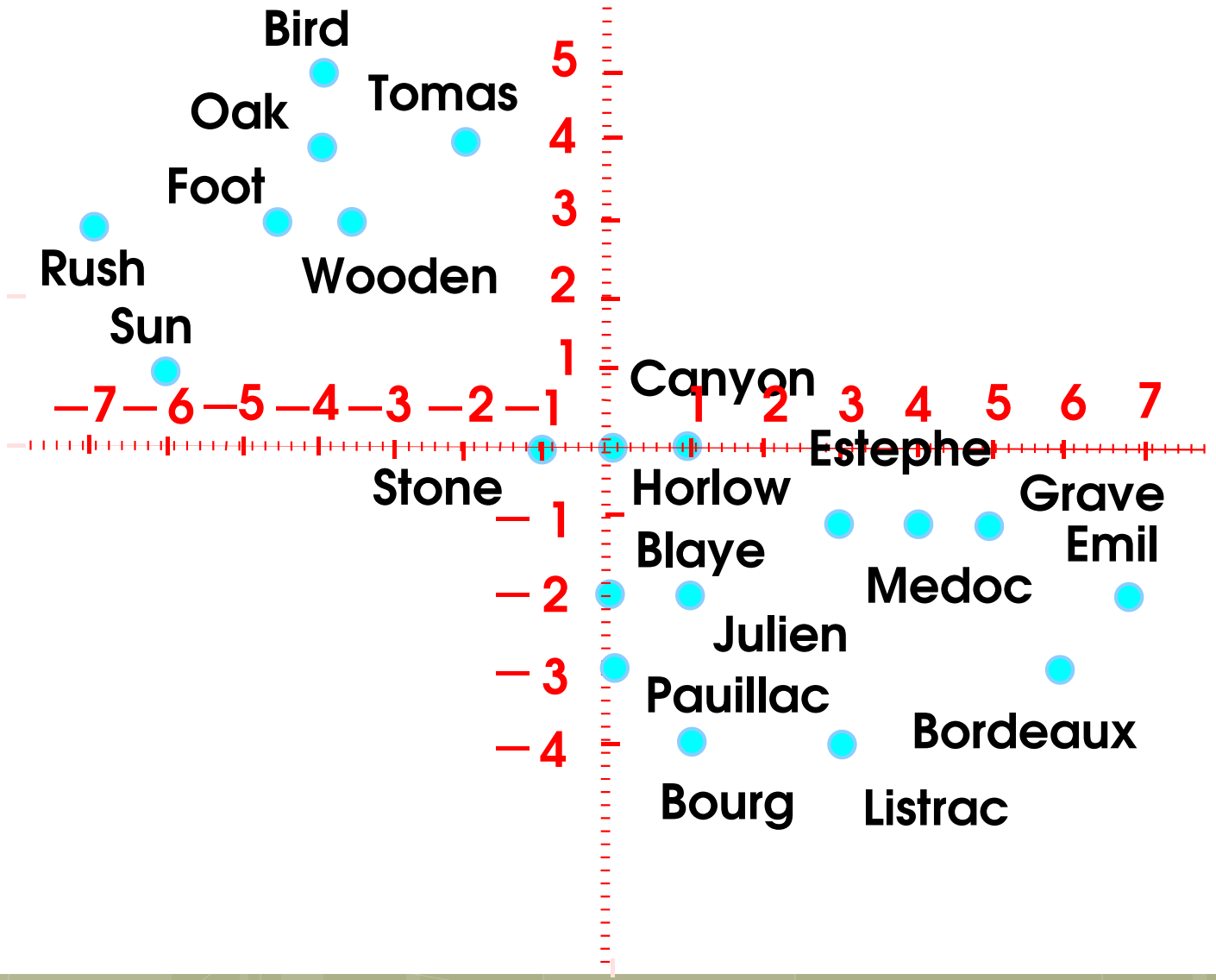


$D^2 = A^2 + B^2$

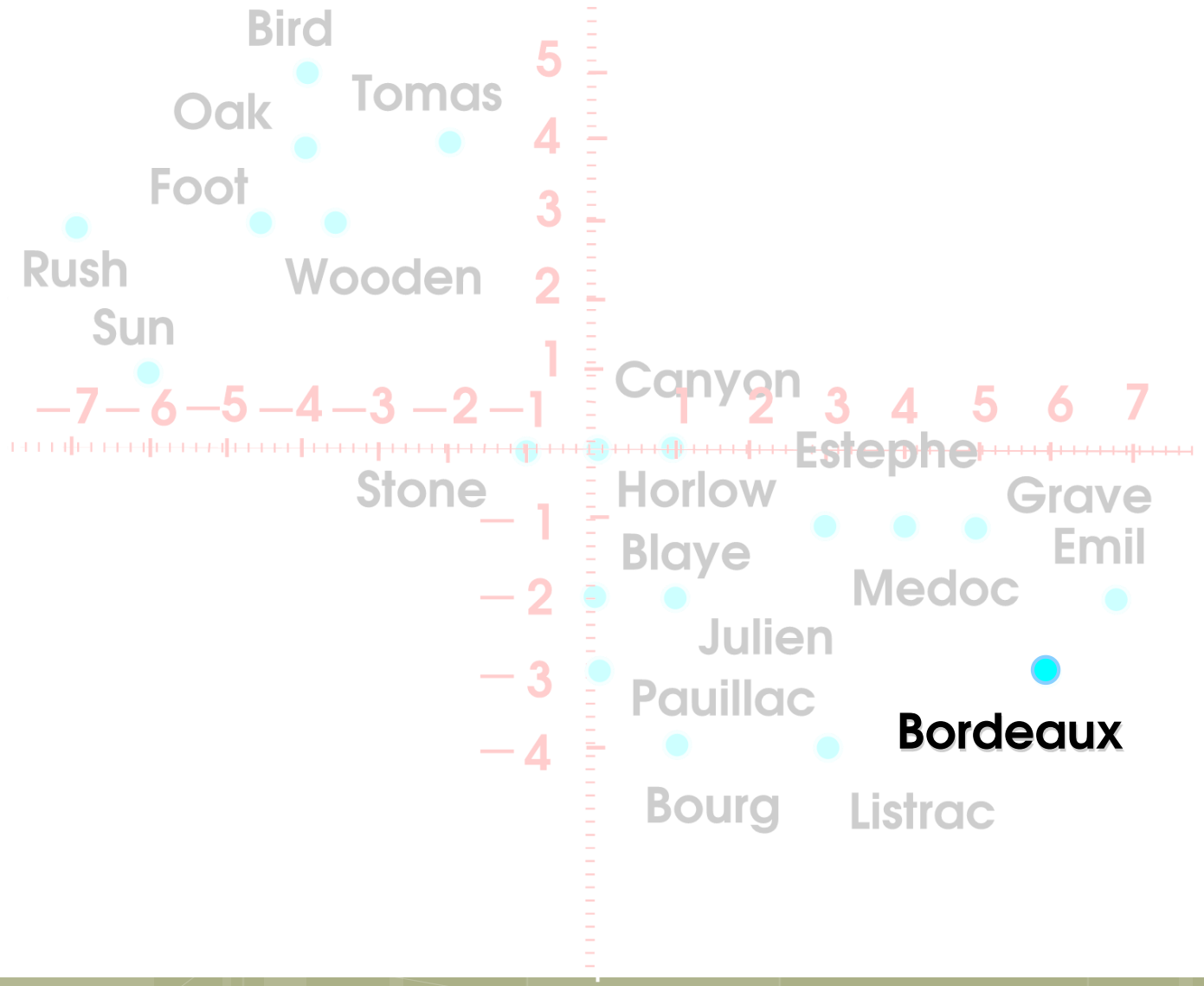
# Pythagorean magic:
## Smallest sum of squared distances same as Largest squared projections (variance)



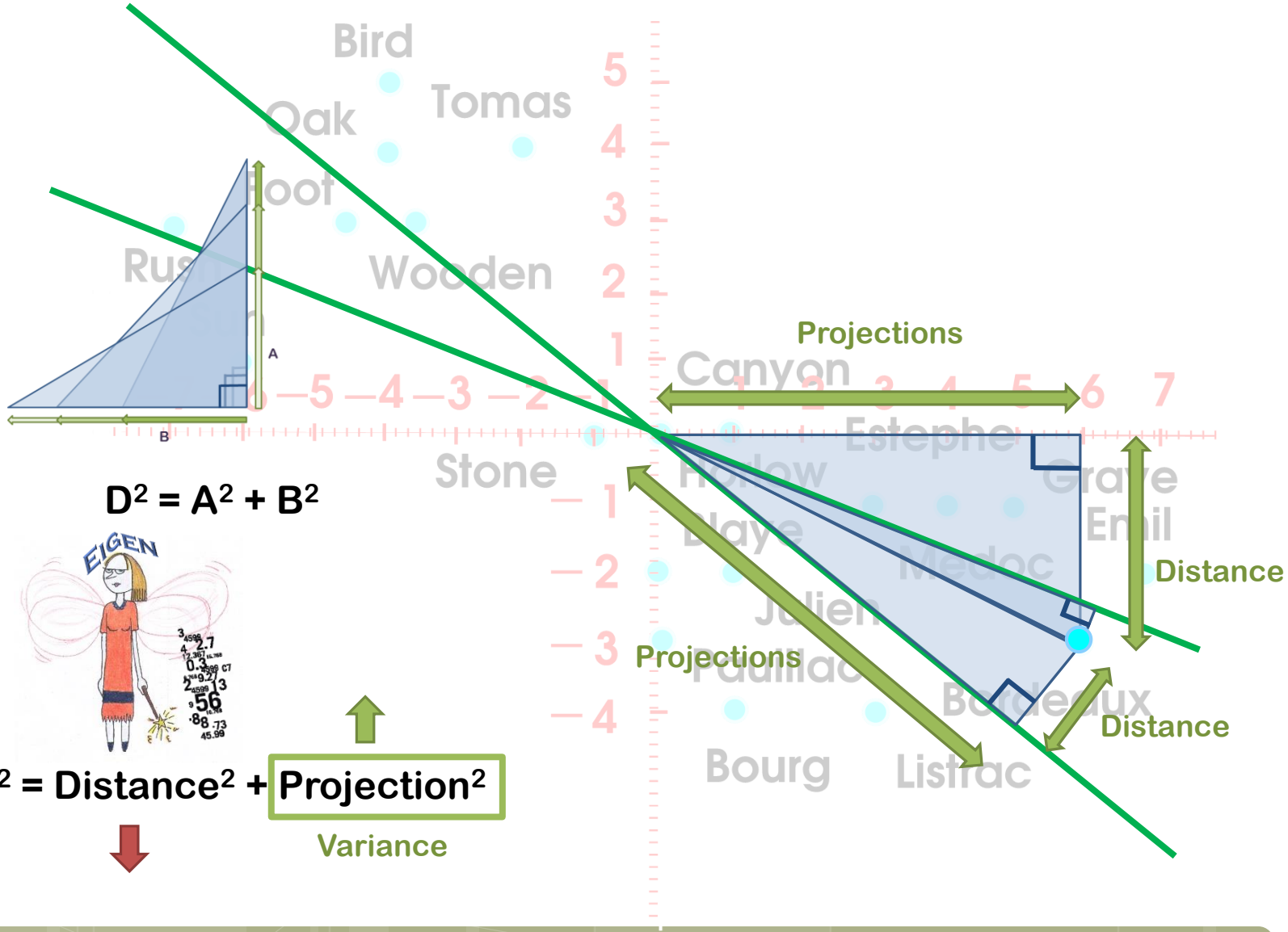
# Back to the picture



# Back to the picture



# Back to the picture



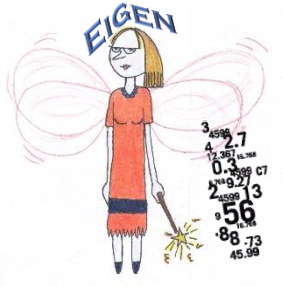

$D^2 = A^2 + B^2$




$D^2 = \text{Distance}^2 + \text{Projection}^2$

Variance

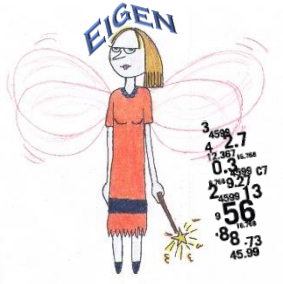
# A PICTURE ... PC1




$D^2 = Distance^2 + Projection^2$

Bird
Tomas
Oak
Foot
Rush
Wooden
Sun
Stone
Canyon
Estephe
Horlow
Grave
Emil
Blaye
Medoc
Julien
Pauillac
Bordeaux
Bourg
Listrac

# A PICTURE ... PC1

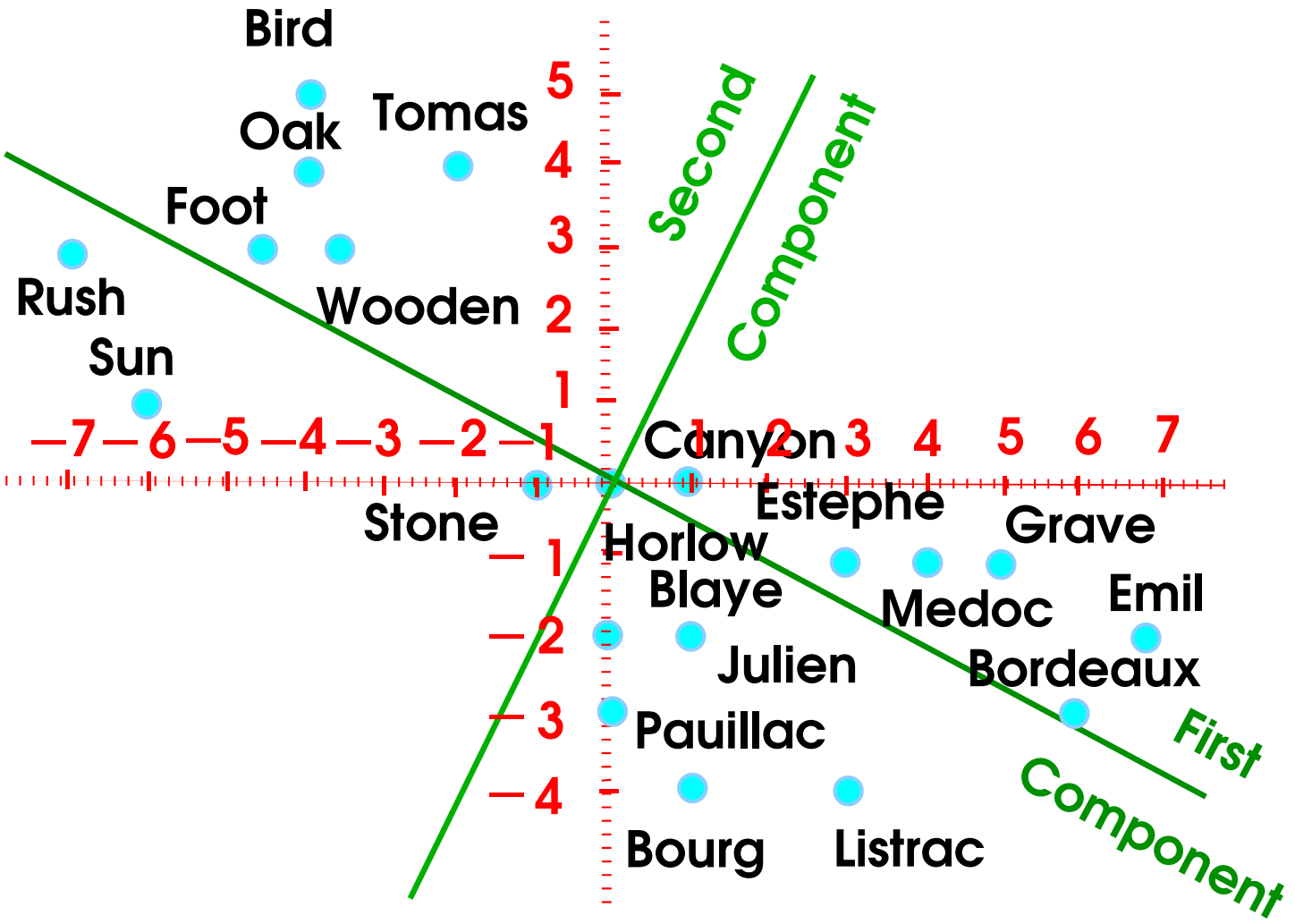



$D^2 = \text{Distance}^2 + \text{Projection}^2$

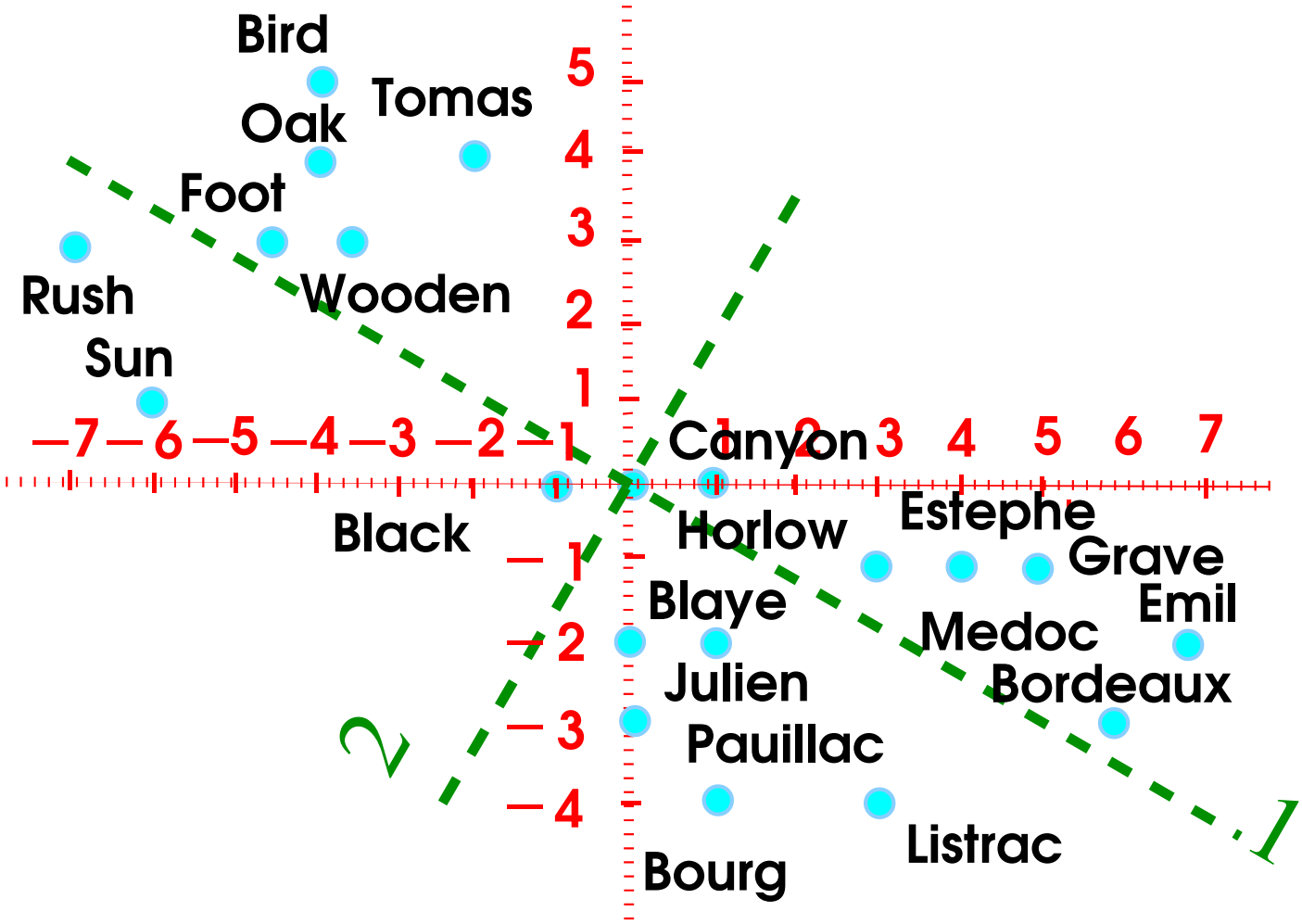
Variance

Bird
Tomas
Oak
Foot
Rush
Wooden
Sun
Canyon
Stone
Horlow
Estephe
Grave
Blaye
Medoc
Emil
Julien
Bordeaux
Pauillac
Listrac
Bourg

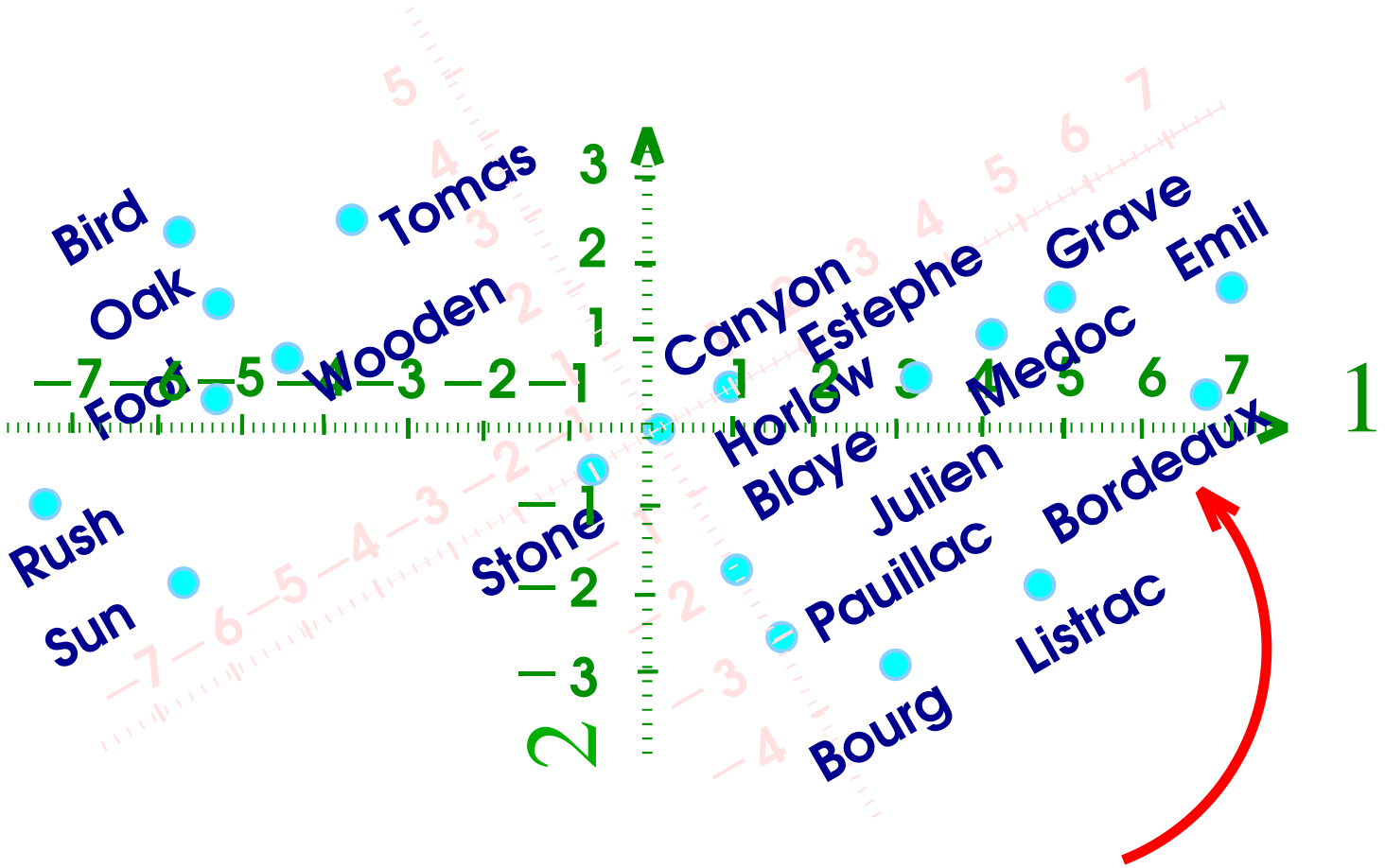
# A PICTURE ... PC1



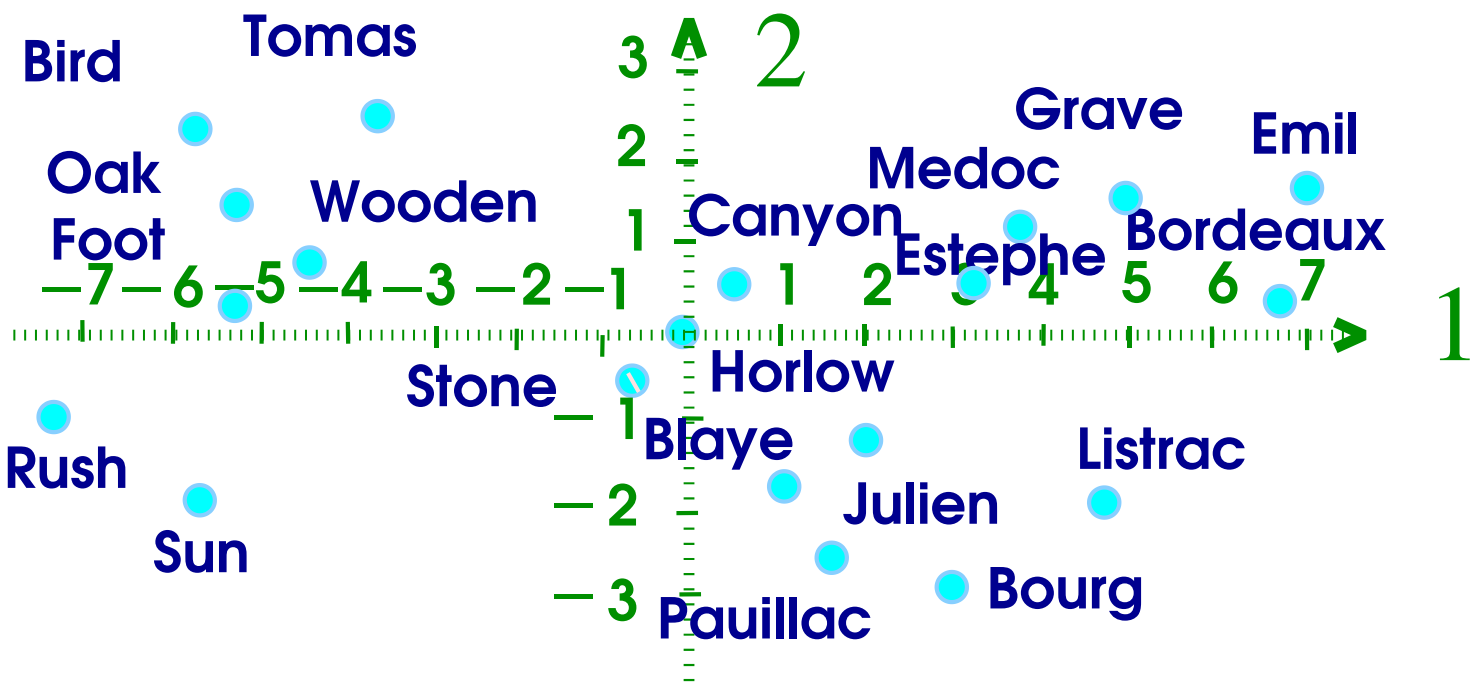
# A PICTURE ... PC1& 2



# A picture ... PC1



# A PICTURE ... PC1 & 2



# This is all there is to PCA!

# How to interpret the dimensions?

# 20 Wines: All data

Here are the data

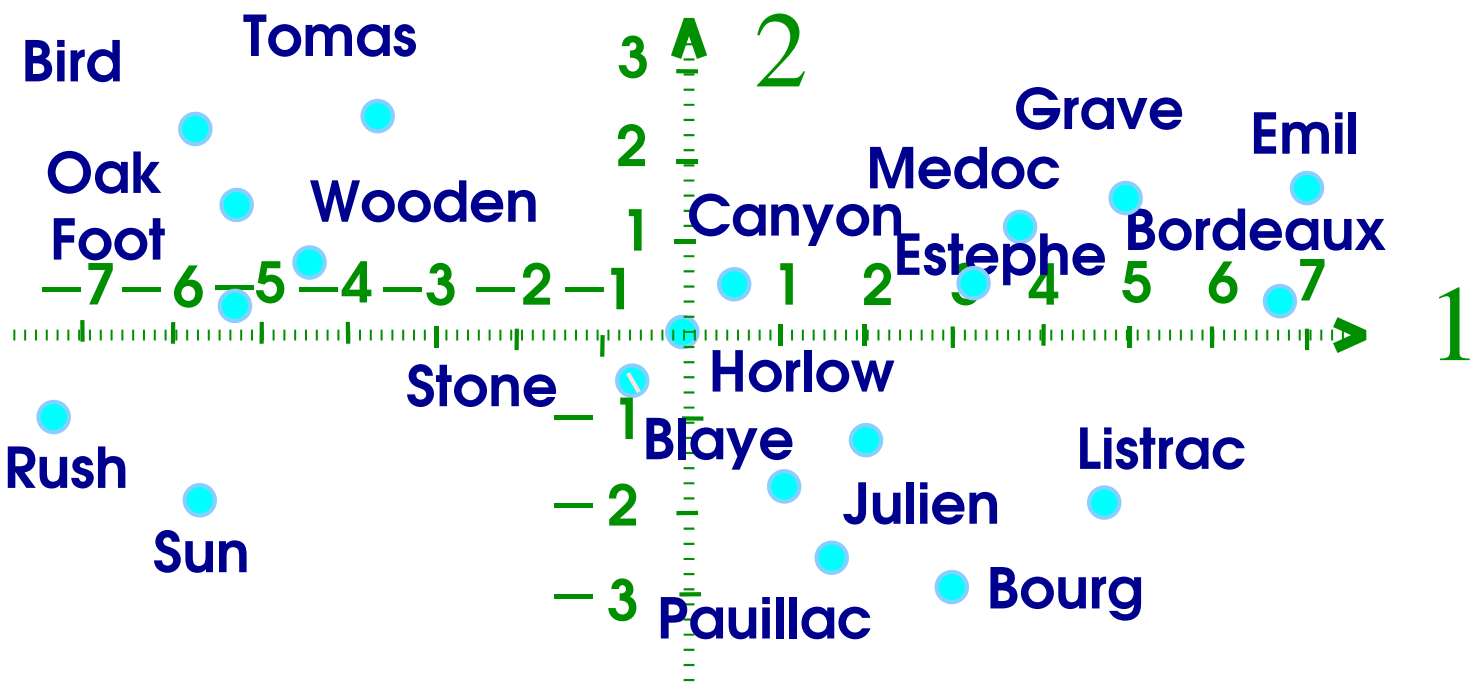
In case you had forgotten. USA vs France

| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Lisrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Hordew | 8. St. Estèphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emillion | 20. Tomsello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Lisrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Hordew | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14 Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| Y: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| W: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |

# A PICTURE ... PC 1 & 2

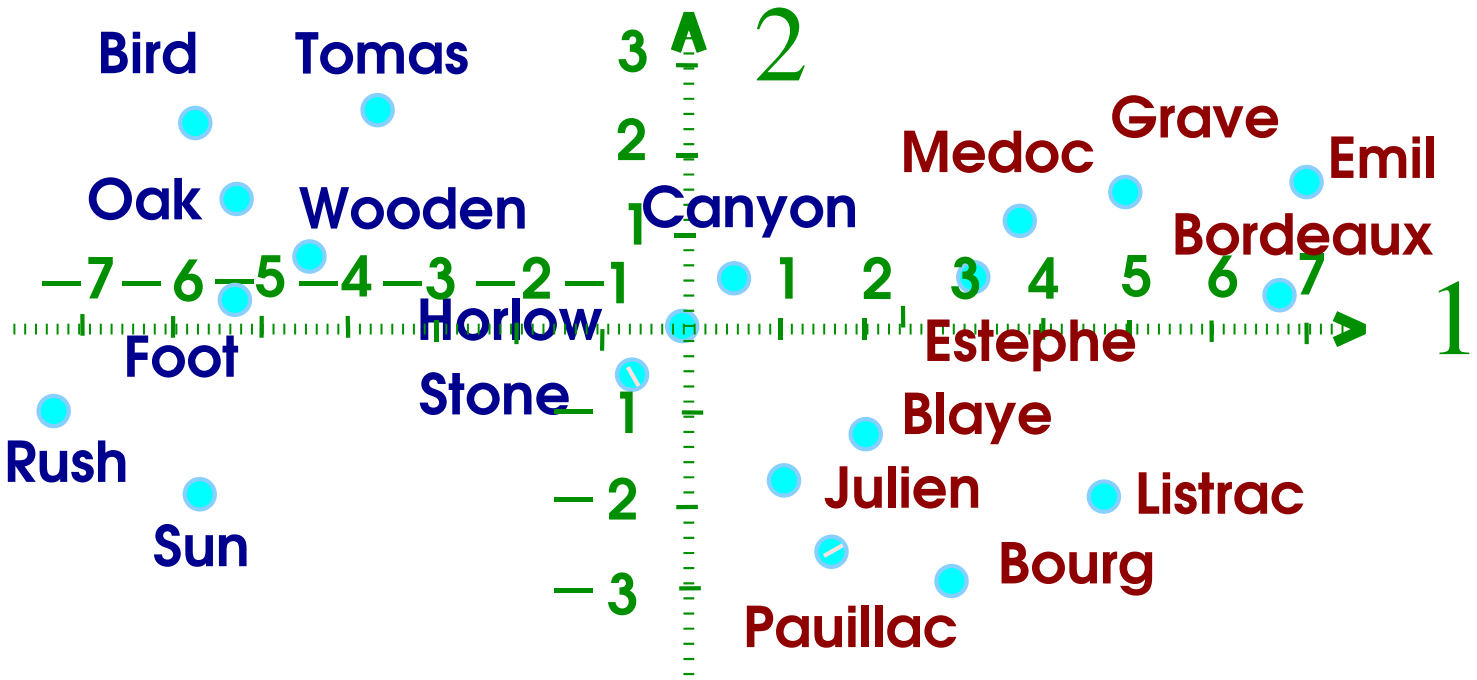
How to interpret the dimensions?

Use COLORS

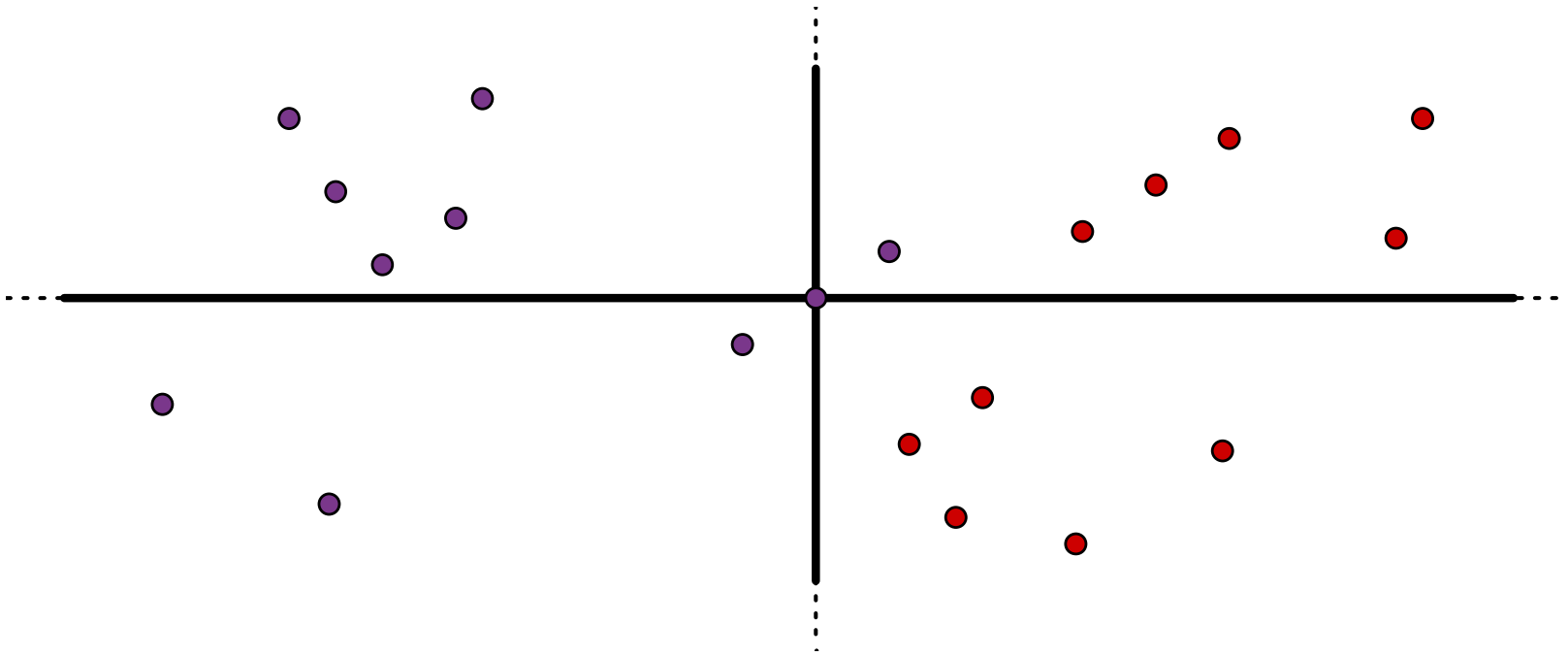


# A PICTURE ... PC1

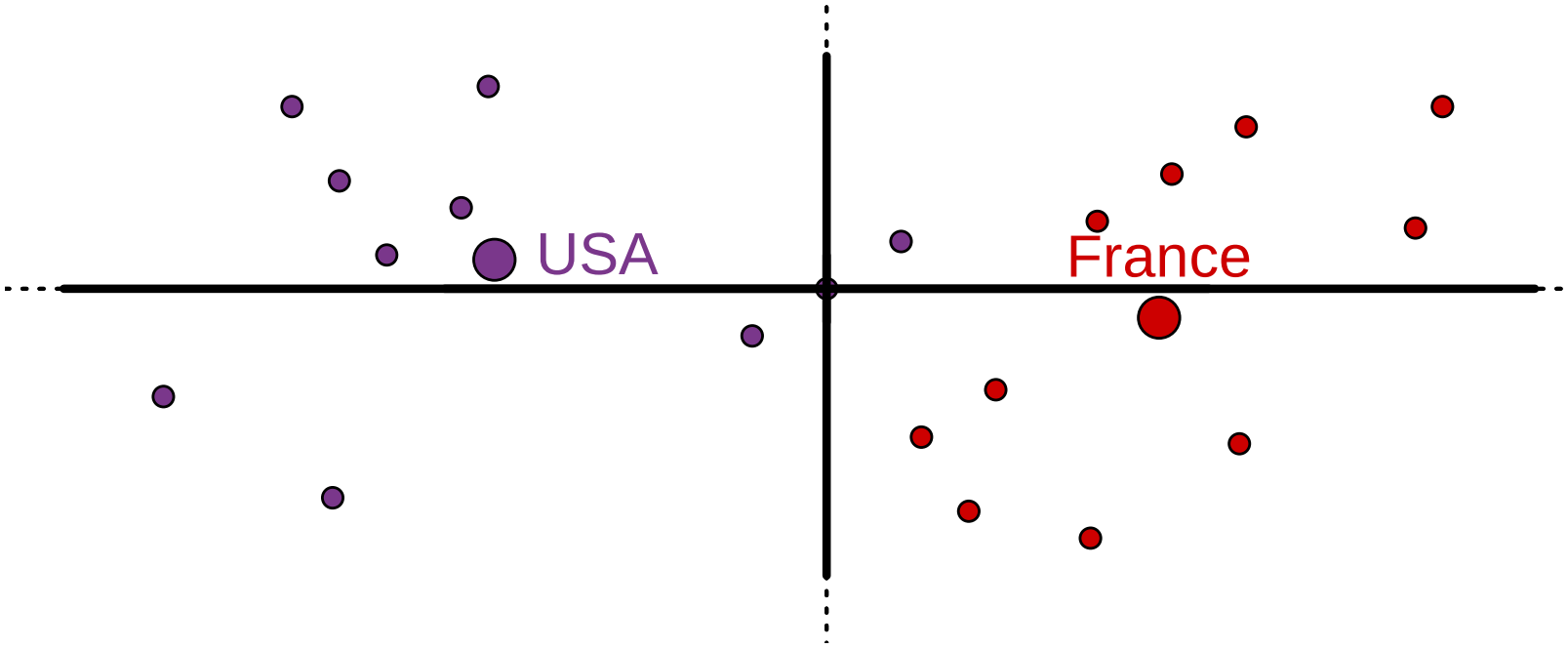
## USA versus France



# With Group Means as Labels



# With Group Means as Labels



# 20 Wines: All data

Here are the data

In case you had forgotten.

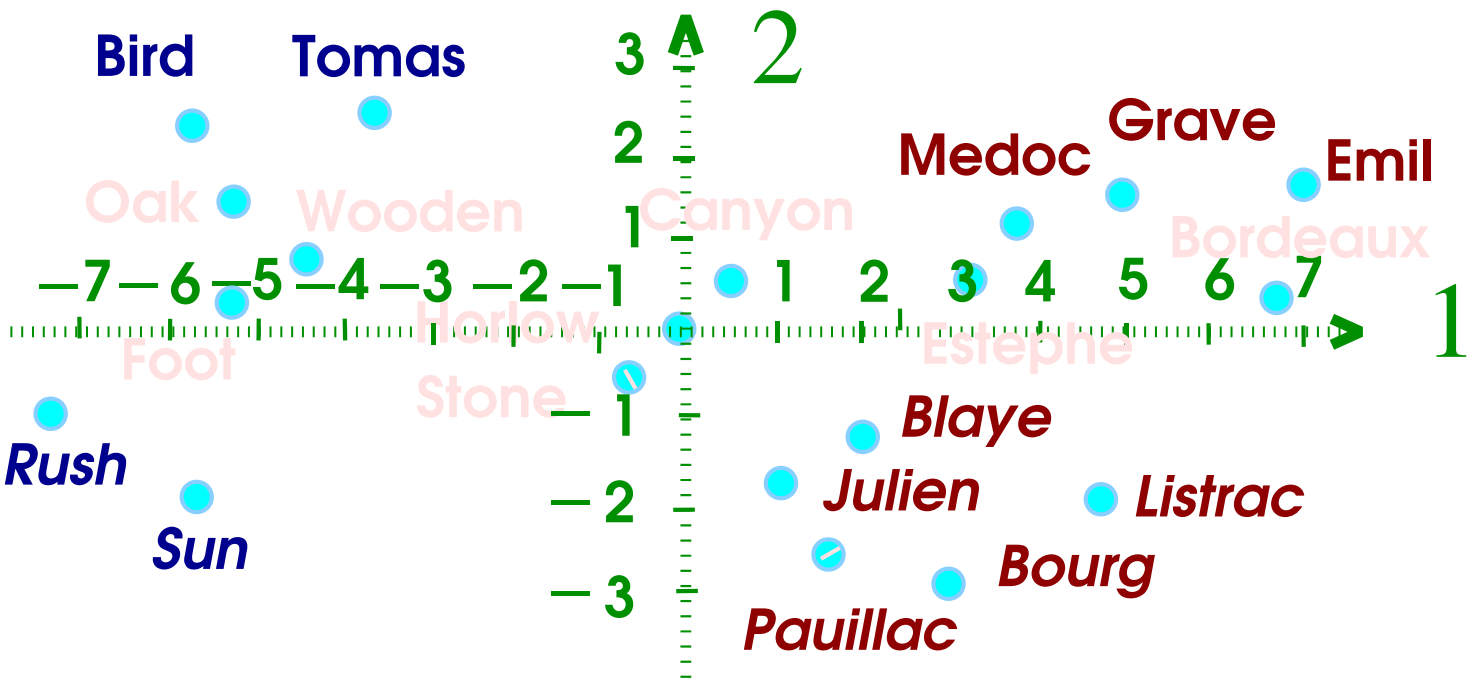
| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Listrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Hordow | 8. St. Estèphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emillion | 20. Tomasello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Listrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Hordow | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14. Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| F: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| W: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |
| Fruity | | | F | | F | | | | | F | F | | | F | F | F | | | | |
| Woody | | | | | | | | | | | | W | W | | | | | W | W | W |
| Acid | 2 | 5 | 1 | 1 | 9 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 9 | 14 | 2 | 1 | 2 | 1 | 1 |
| Bitter | 8 | 3 | 16 | 3 | 11 | 1 | 1 | 9 | 1 | 8 | 3 | 2 | 9 | 12 | 10 | 2 | 1 | 10 | 7 | 3 |

# Some wines are fruity, some are woody

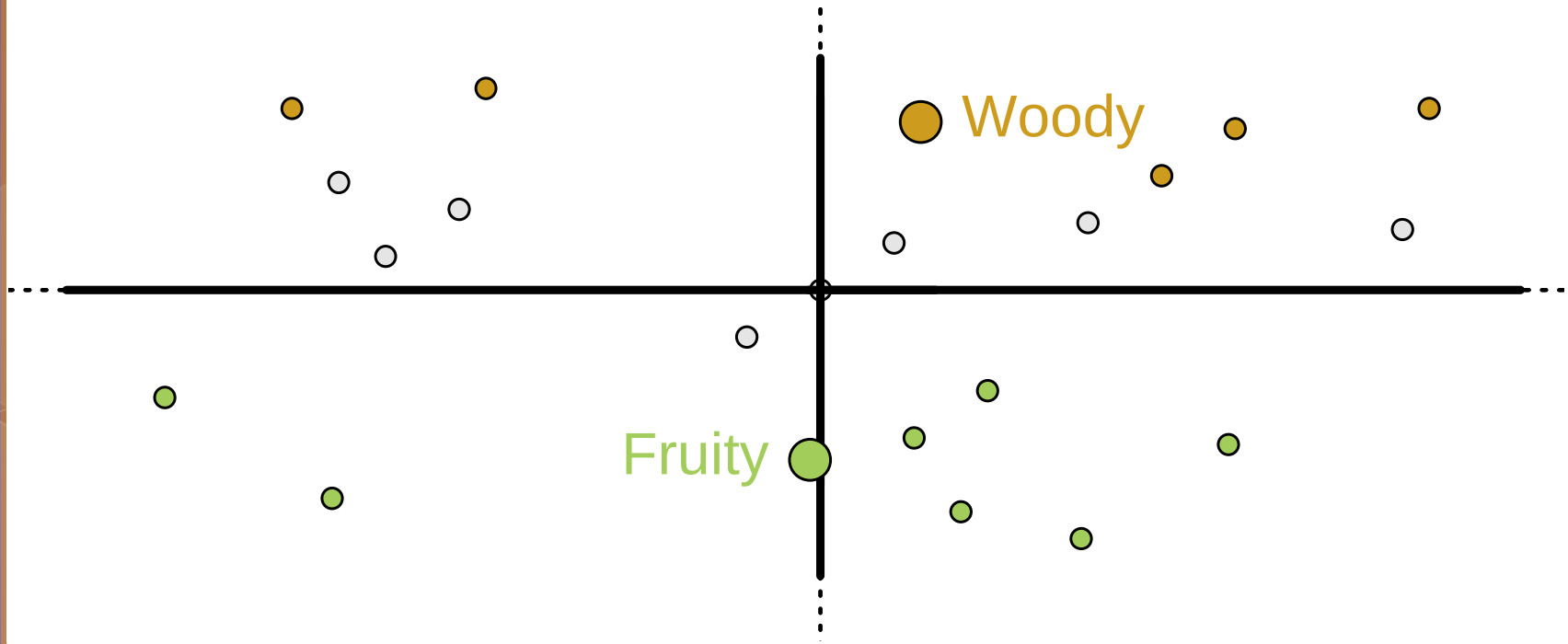
# A PICTURE ... PC2

## SOME WINES ARE FRUITY, SOME ARE WOODY

### Woody versus *Fruity*



# Some wines are fruity, some are woody

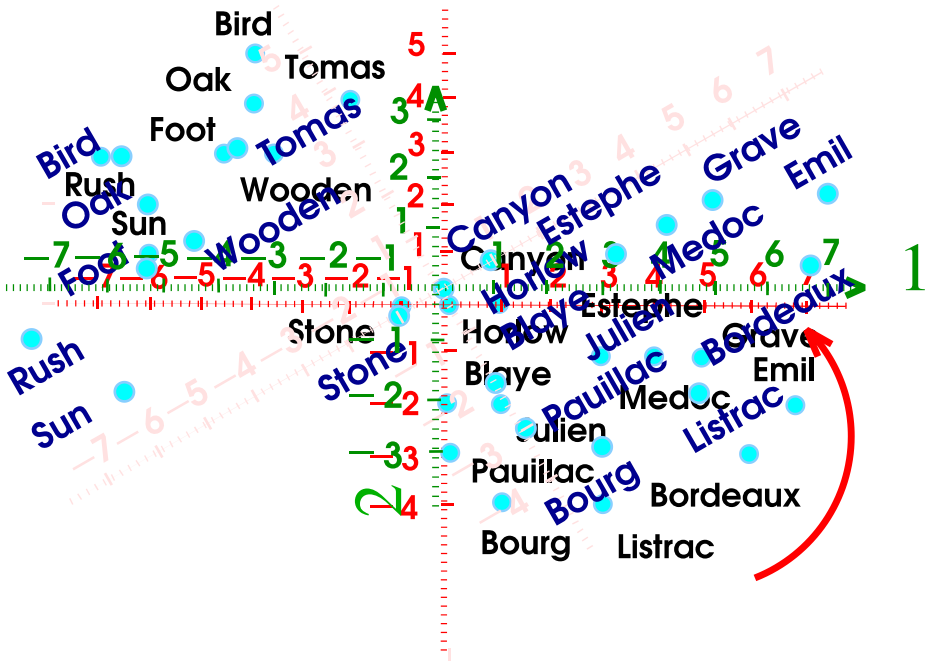


# A BIT MORE FORMAL

# Vocabulary and some Greek

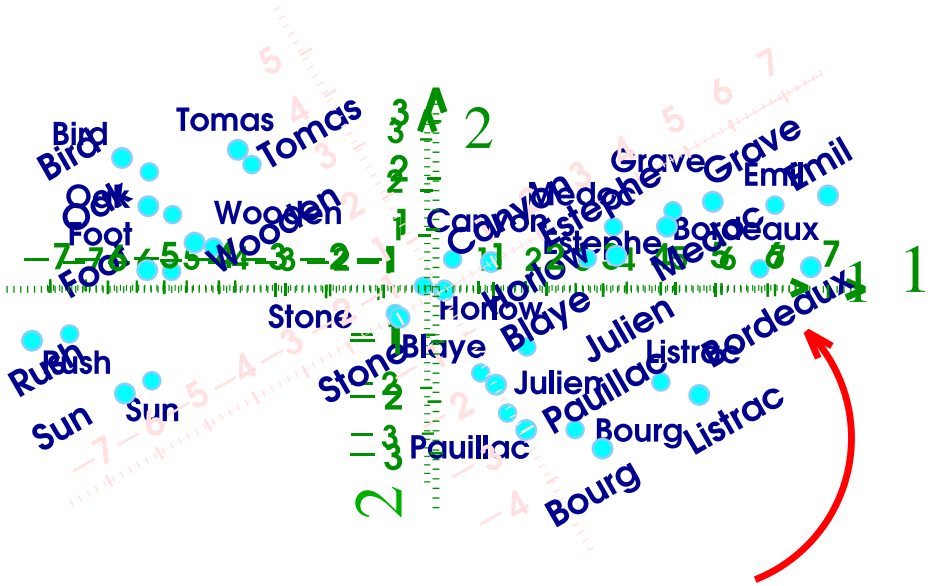
# Factor scores

| | Y | W | y | w |
|---|---|---|---|---|
| Bordeaux | 3 | 14 | −3 | 6 |
| Black Stone | 6 | 7 | 0 | −1 |
| Listrac | 2 | 11 | −4 | 3 |
| Canyon Creek | 6 | 9 | 0 | 1 |
| Côtes de Bourg | 2 | 9 | −4 | 1 |
| Foot Hill | 9 | 4 | 3 | −4 |
| Horlow | 6 | 8 | 0 | 0 |
| St. Estphe | 5 | 11 | −1 | 3 |
| Wooden Hill | 9 | 5 | 3 | −3 |
| Côtes de Blaye | 4 | 8 | −2 | 0 |
| Sun Set | 7 | 2 | 1 | −6 |
| Black Bird | 11 | 4 | 5 | −4 |
| Médoc | 5 | 12 | −1 | 4 |
| St Julien | 4 | 9 | −2 | 1 |
| Pauillac | 3 | 8 | −3 | 0 |
| Gold Rush | 9 | 1 | 3 | −7 |
| Oak Ville | 10 | 4 | 4 | −4 |
| Grave | 5 | 13 | −1 | 5 |
| St Emilion | 4 | 15 | −2 | 7 |
| Tomasello | 10 | 6 | 4 | −2 |
| Σ | 120 | 160 | 0 | 0 |



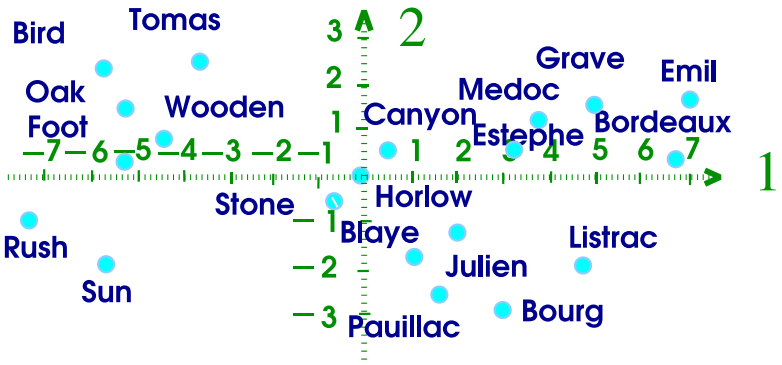
# CALL THE COORDINATES: FACTOR SCORES

| | Y | W | y | w | $F_1$ | $F_2$ |
|---|---|---|---|---|---|---|
| Bordeaux | 3 | 14 | −3 | 6 | 6.67 | 0.69 |
| Black Stone | 6 | 7 | 0 | −1 | −0.84 | −0.54 |
| Listrac | 2 | 11 | −4 | 3 | 4.68 | −1.76 |
| Canyon Creek | 6 | 9 | 0 | 1 | 0.84 | 0.54 |
| Côtes de Bourg | 2 | 9 | −4 | 1 | 2.99 | −2.84 |
| Foot Hill | 9 | 4 | 3 | −4 | −4.99 | 0.38 |
| Horlow | 6 | 8 | 0 | 0 | 0.00 | 0.00 |
| St. Estphe | 5 | 11 | −1 | 3 | 3.07 | 0.77 |
| Wooden Hill | 9 | 5 | 3 | −3 | −4.14 | 0.92 |
| Côtes de Blaye | 4 | 8 | −2 | 0 | 1.07 | −1.69 |
| Sun Set | 7 | 2 | 1 | −6 | −5.60 | −2.38 |
| Black Bird | 11 | 4 | 5 | −4 | −6.06 | 2.07 |
| Médoc | 5 | 12 | −1 | 4 | 3.91 | 1.30 |
| St Julien | 4 | 9 | −2 | 1 | 1.92 | −1.15 |
| Pauillac | 3 | 8 | −3 | 0 | 1.61 | −2.53 |
| Gold Rush | 9 | 1 | 3 | −7 | −7.52 | −1.23 |
| Oak Ville | 10 | 4 | 4 | −4 | −5.52 | 1.23 |
| Grave | 5 | 13 | −1 | 5 | 4.76 | 1.84 |
| St Emilion | 4 | 15 | −2 | 7 | 6.98 | 2.07 |
| Tomasello | 10 | 6 | 4 | −2 | −3.83 | 2.30 |
| Σ | 120 | 160 | 0 | 0 | 0 | 0 |



# Call Sum of squared factor scores: Eigenvalue

| | $Y$ | $W$ | $y$ | $w$ | $F_1$ | $F_2$ | $F_1^2$ | $F_2^2$ |
|---|---|---|---|---|---|---|---|---|
| Bordeaux | 3 | 14 | −3 | 6 | 6.67 | 0.69 | 44.52 | 0.48 |
| Black Stone | 6 | 7 | 0 | −1 | −0.84 | −0.54 | 0.71 | 0.29 |
| Listrac | 2 | 11 | −4 | 3 | 4.68 | −1.76 | 21.89 | 3.11 |
| Canyon Creek | 6 | 9 | 0 | 1 | 0.84 | 0.54 | 0.71 | 0.29 |
| Côtes de Bourg | 2 | 9 | −4 | 1 | 2.99 | −2.84 | 8.95 | 8.05 |
| Foot Hill | 9 | 4 | 3 | −4 | −4.99 | 0.38 | 24.85 | 0.15 |
| Horlow | 6 | 8 | 0 | 0 | 0.00 | 0.00 | 0 | 0.00 |
| St. Estphe | 5 | 11 | −1 | 3 | 3.07 | 0.77 | 9.41 | 0.59 |
| Wooden Hill | 9 | 5 | 3 | −3 | −4.14 | 0.92 | 17.15 | 0.85 |
| Côtes de Blaye | 4 | 8 | −2 | 0 | 1.07 | −1.69 | 1.15 | 2.85 |
| Sun Set | 7 | 2 | 1 | −6 | −5.60 | −2.38 | 31.35 | 5.65 |
| Black Bird | 11 | 4 | 5 | −4 | −6.06 | 2.07 | 36.71 | 4.29 |
| Médoc | 5 | 12 | −1 | 4 | 3.91 | 1.30 | 15.30 | 1.70 |
| St Julien | 4 | 9 | −2 | 1 | 1.92 | −1.15 | 3.68 | 1.32 |
| Pauillac | 3 | 8 | −3 | 0 | 1.61 | −2.53 | 2.59 | 6.41 |
| Gold Rush | 9 | 1 | 3 | −7 | −7.52 | −1.23 | 56.49 | 1.51 |
| Oak Ville | 10 | 4 | 4 | −4 | −5.52 | 1.23 | 30.49 | 1.51 |
| Grave | 5 | 13 | −1 | 5 | 4.76 | 1.84 | 22.61 | 3.39 |
| St Emilion | 4 | 15 | −2 | 7 | 6.98 | 2.07 | 48.71 | 4.29 |
| Tomasello | 10 | 6 | 4 | −2 | −3.83 | 2.30 | 14.71 | 5.29 |
| $\Sigma$ | 120 | 160 | 0 | 0 | 0 | 0 | 392 | 52 |
| | | | | | | | $\lambda_1$ | $\lambda_2$ |



# Explained variance?

# The revenge of Pythagoras

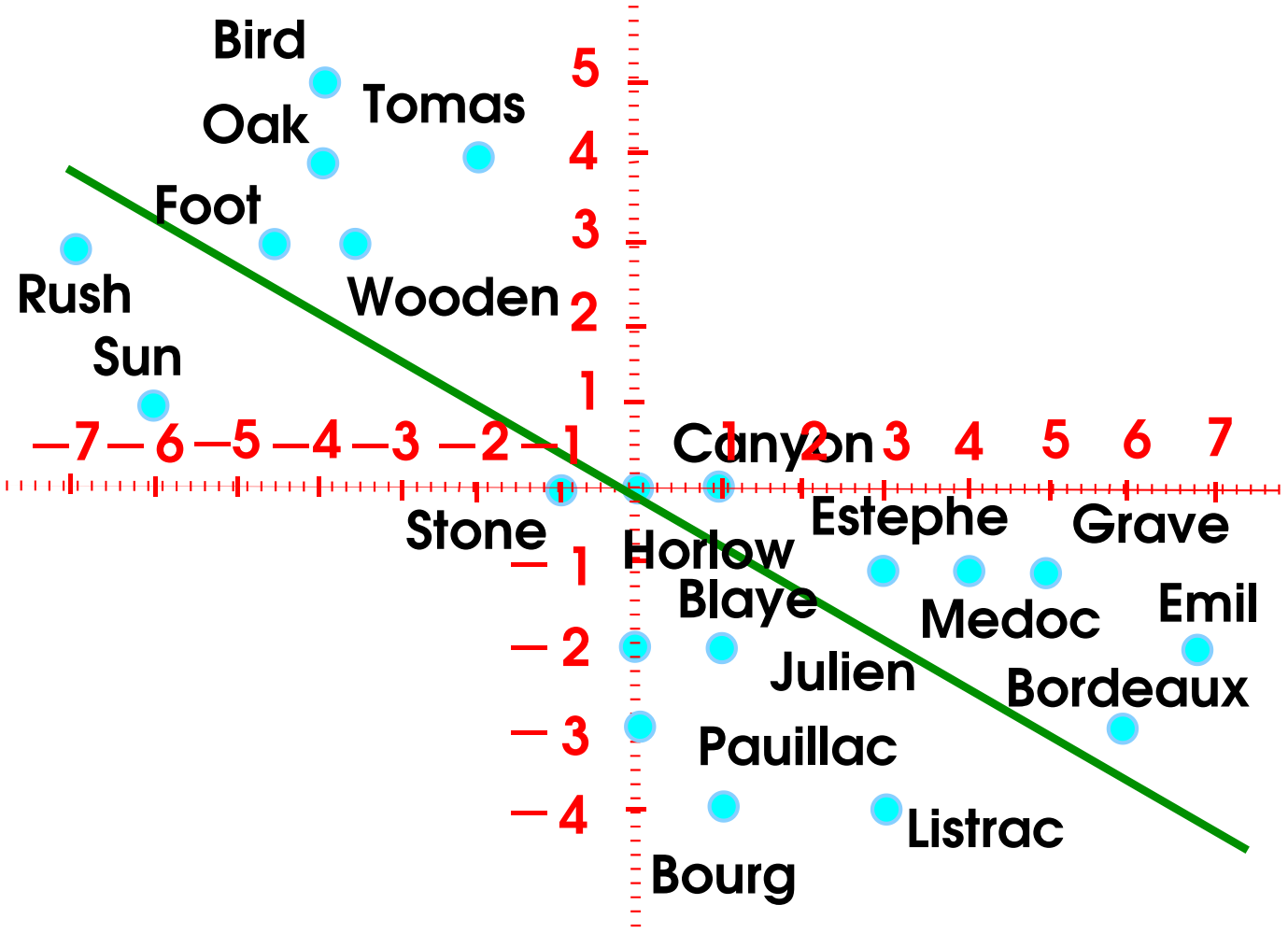
# The return of the Eigen-Fairy

# Back to the Eigen Fairy

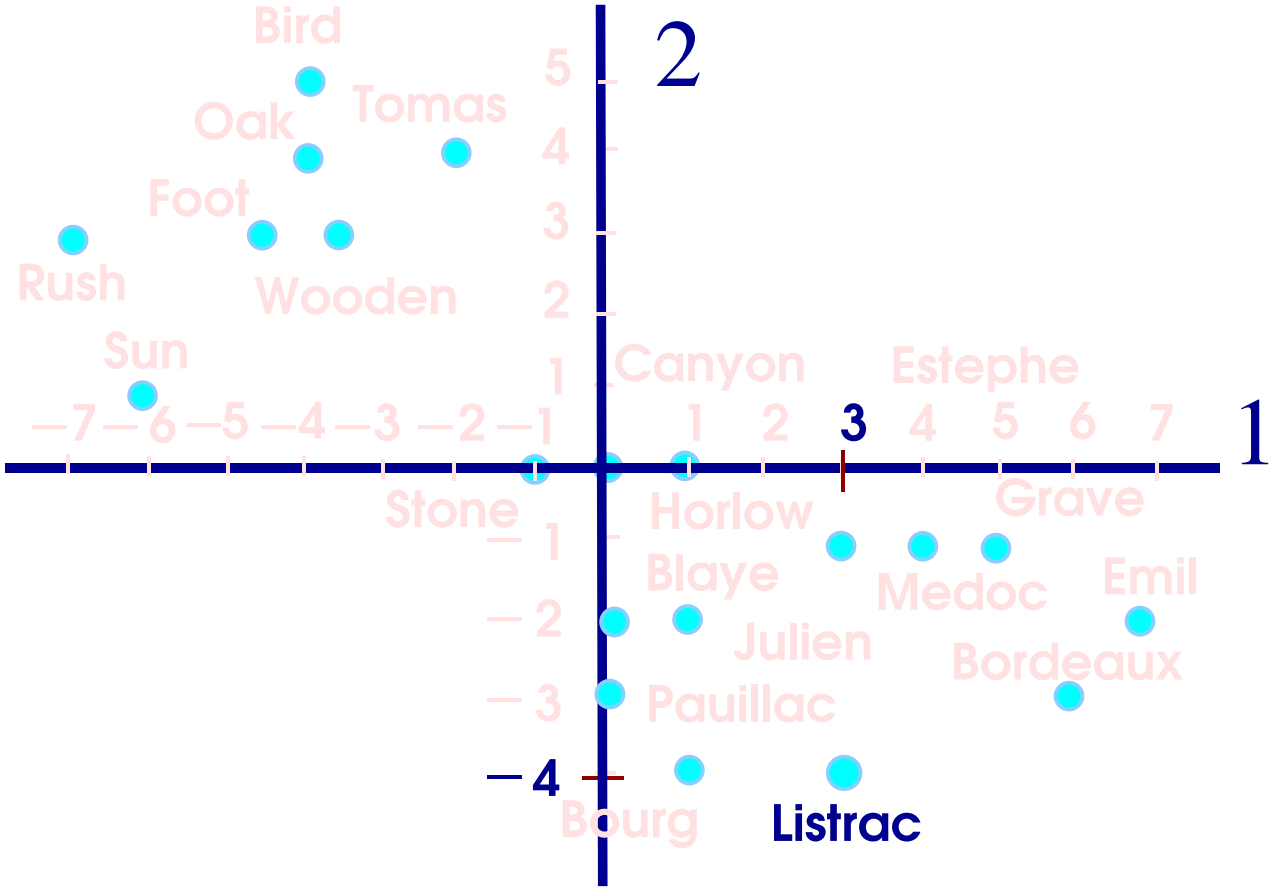

## What did we ask for?

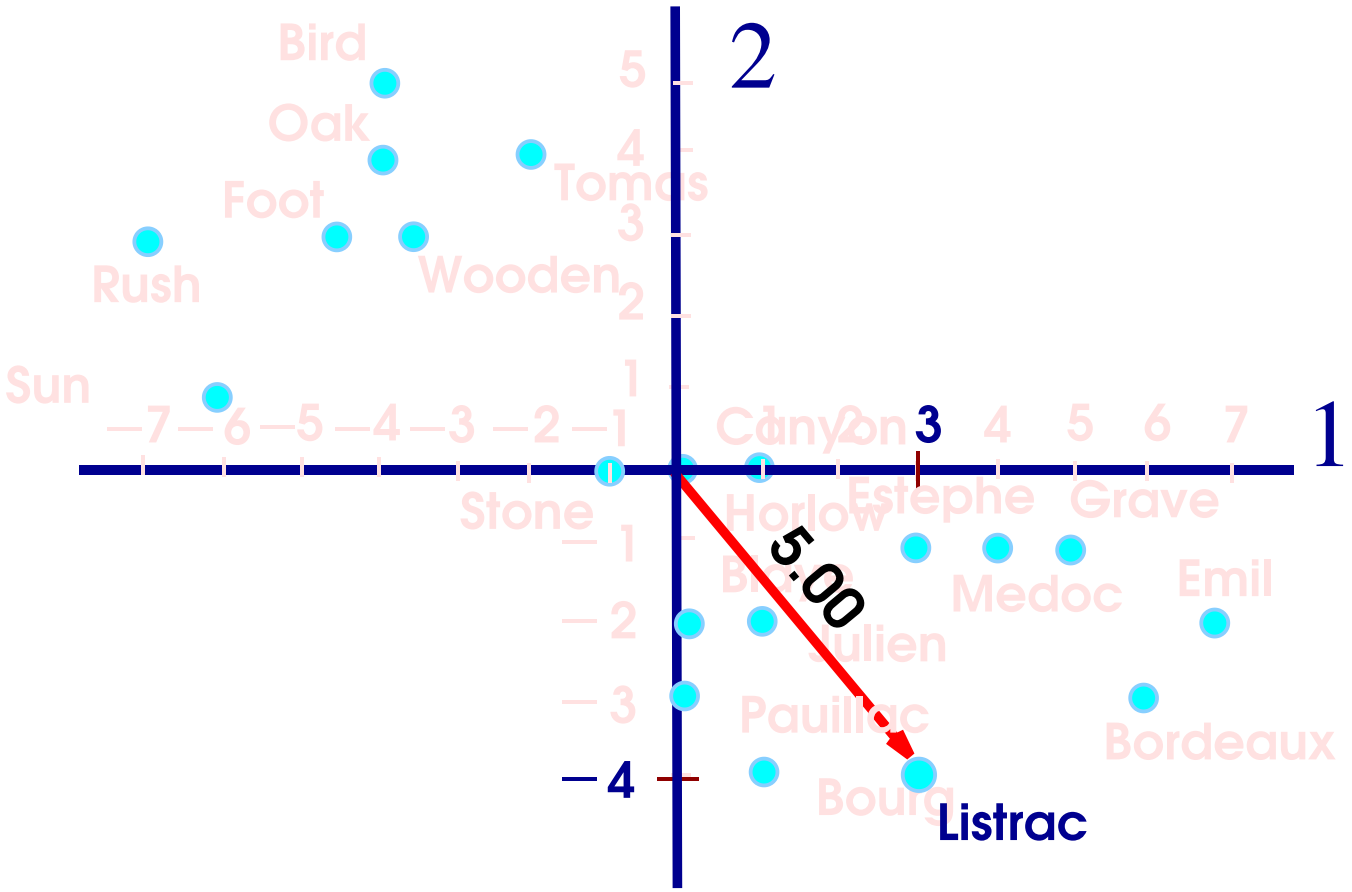
Eigen Fairy, pleeeez
Find the rotation
With smallest
Squared distances

# A PICTURE ... PC1



# The square distances




Bird
Oak
Foot
Tomas
Rush
Wooden
Sun
Canyon
Stone
Horlow
Estephe
Grave
Medoc
Emil
Julien
Pauillac
Bordeaux
Bourg
Listrac
5.00
1
2
3
4
5
6
7
−7
−6
−5
−4
−3
−2
−1
−4

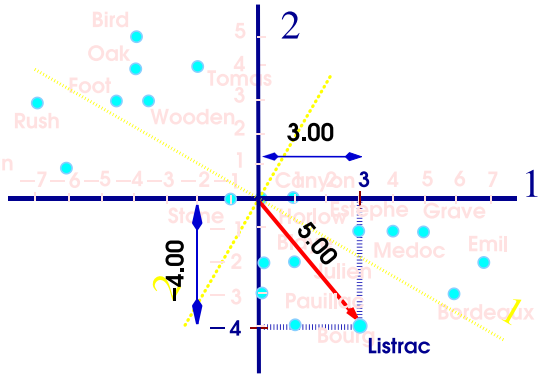
## Remember

# The Theorem!

# Inertia: What do we decompose in PCA?

$$d^2_{\text{Listrac},\mathbf{g}} = w^2_{\text{Listrac},\mathbf{g}} + y^2_{\text{Listrac},\mathbf{g}}$$

$$3^2 + (-4)^2 = 9 + 16 = 25$$

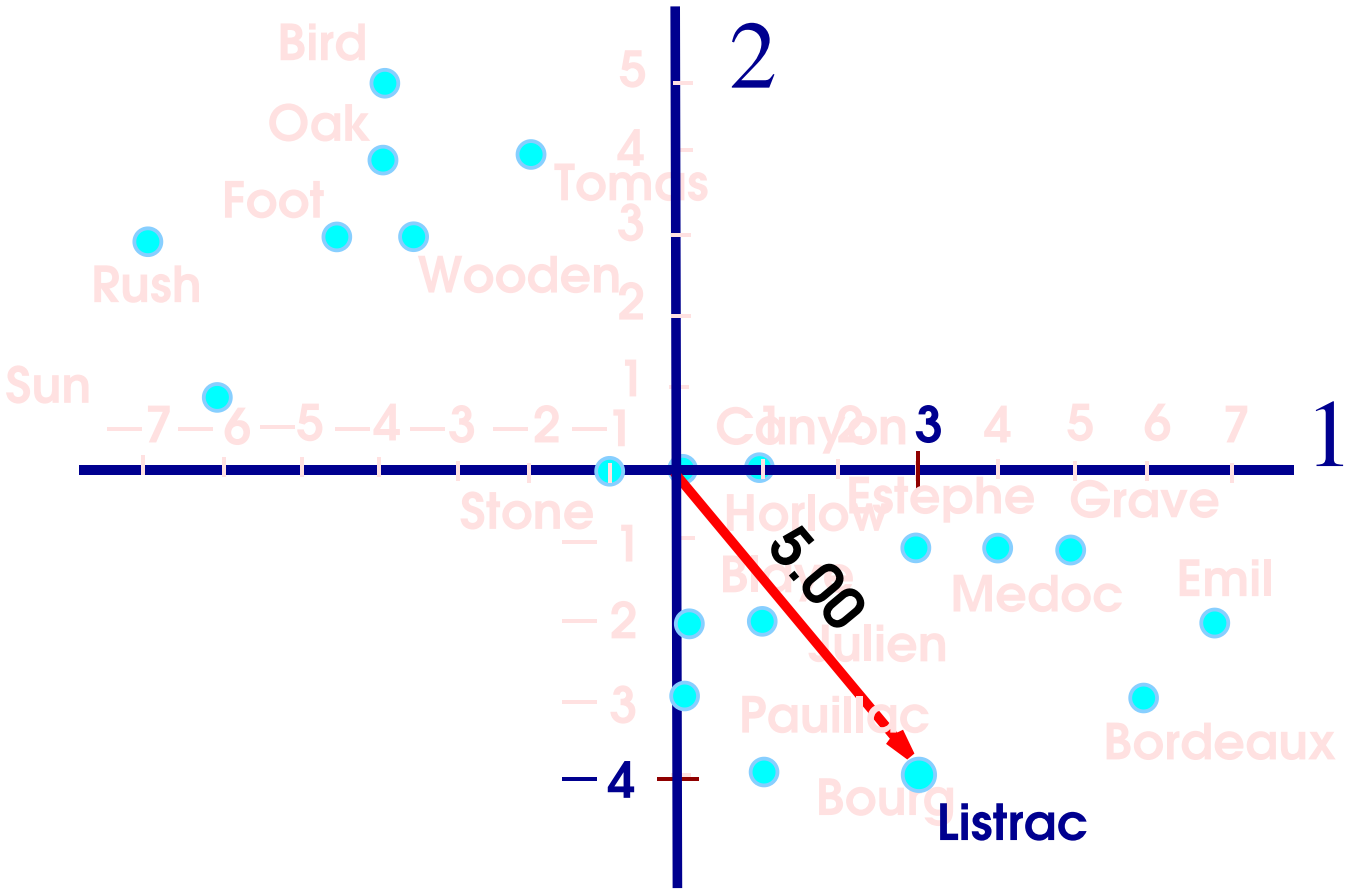



# Inertia: Sum of squared distances

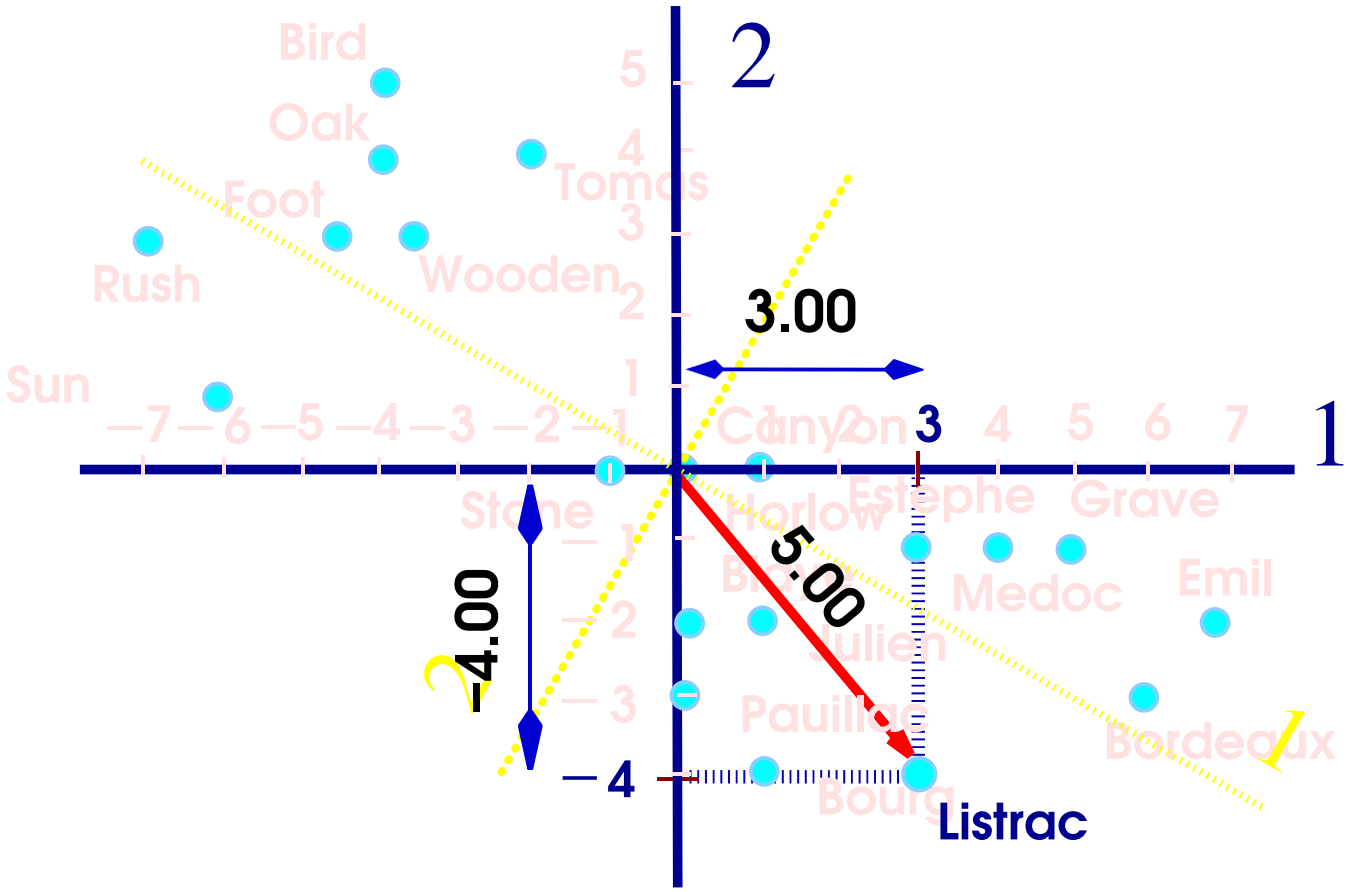
$$\mathcal{I} = \sum_{i}^{I} d^2_{i,\mathbf{g}}$$

# INERTIA FROM THE VARIABLES

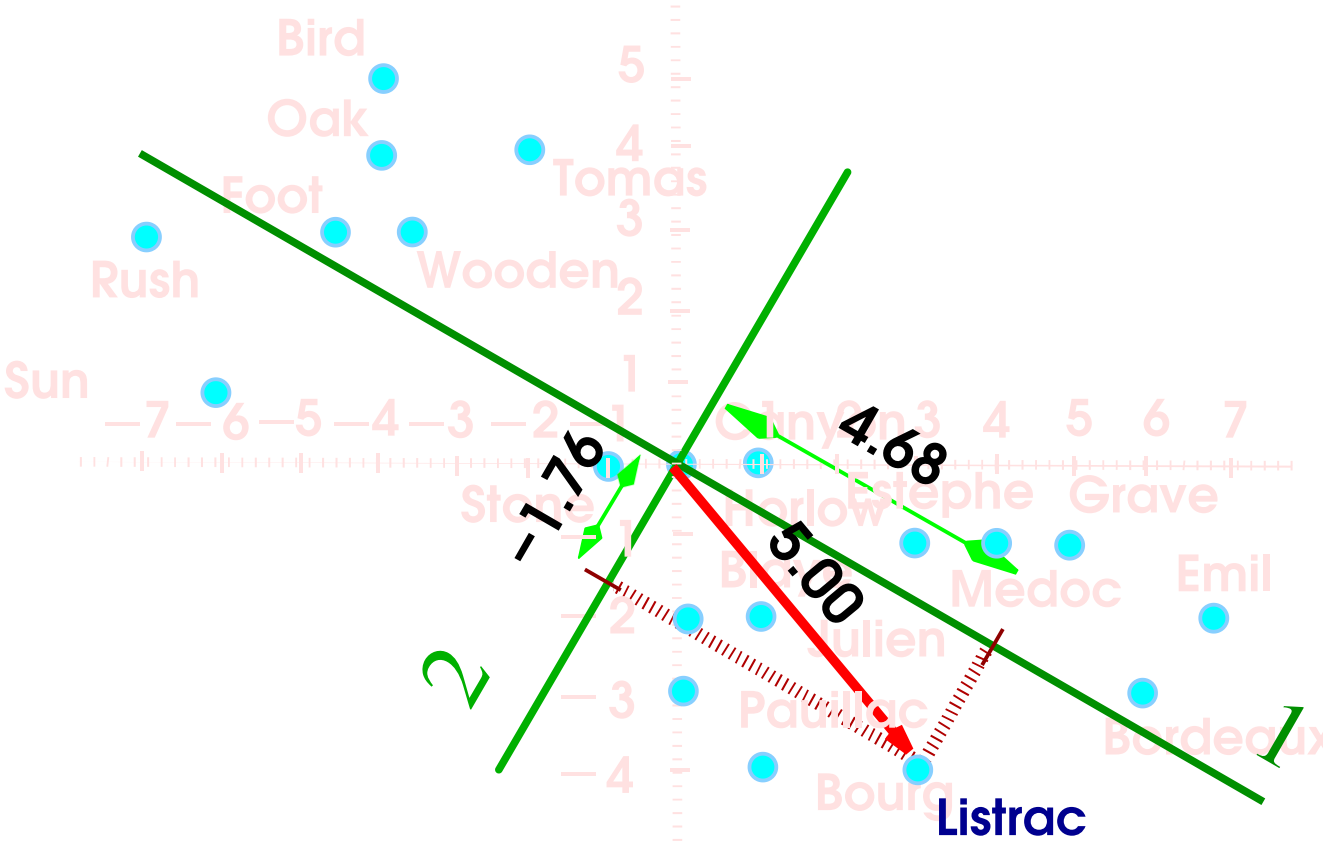
$$\mathcal{I} = \sum_{i}^{I} d^2_{i,\mathbf{g}} = \sum_{i}^{I} \left( w^2_{i,\mathbf{g}} + y^2_{i,\mathbf{g}} \right)$$


Bird
Oak
Foot
Tomas
Rush
Wooden
Sun
Canyon
Horlow
Estephe
Grave
Stone
Medoc
Emil
Julien
Pauillac
Bordeaux
Bourg
Listrac
5.00
1
2
3
4
5
6
7
−7
−6
−5
−4
−3
−2
−1
1
2
3
4
5
−1
−2
−3
−4

# Distance from Variables



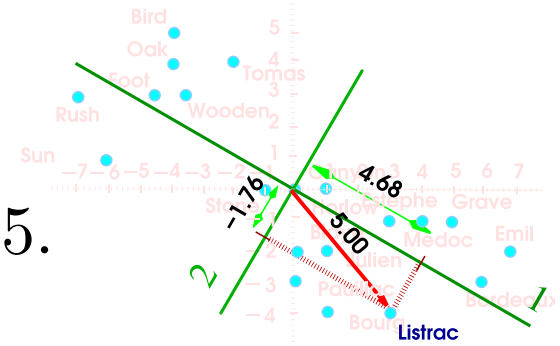
# Distance from components



# INERTIA FROM THE COMPONENTS

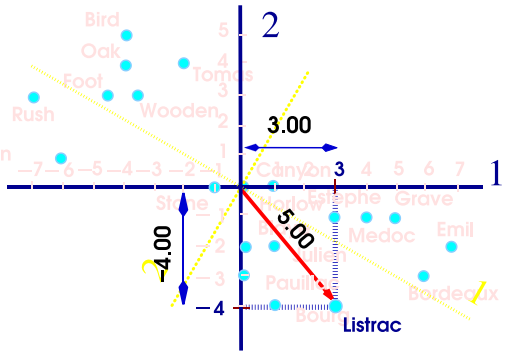
**Squared distance from "Listrac" to center of the space From components**

$$4.68^2 + (-1.76)^2 = 21.89 + 3.11 = 25.$$



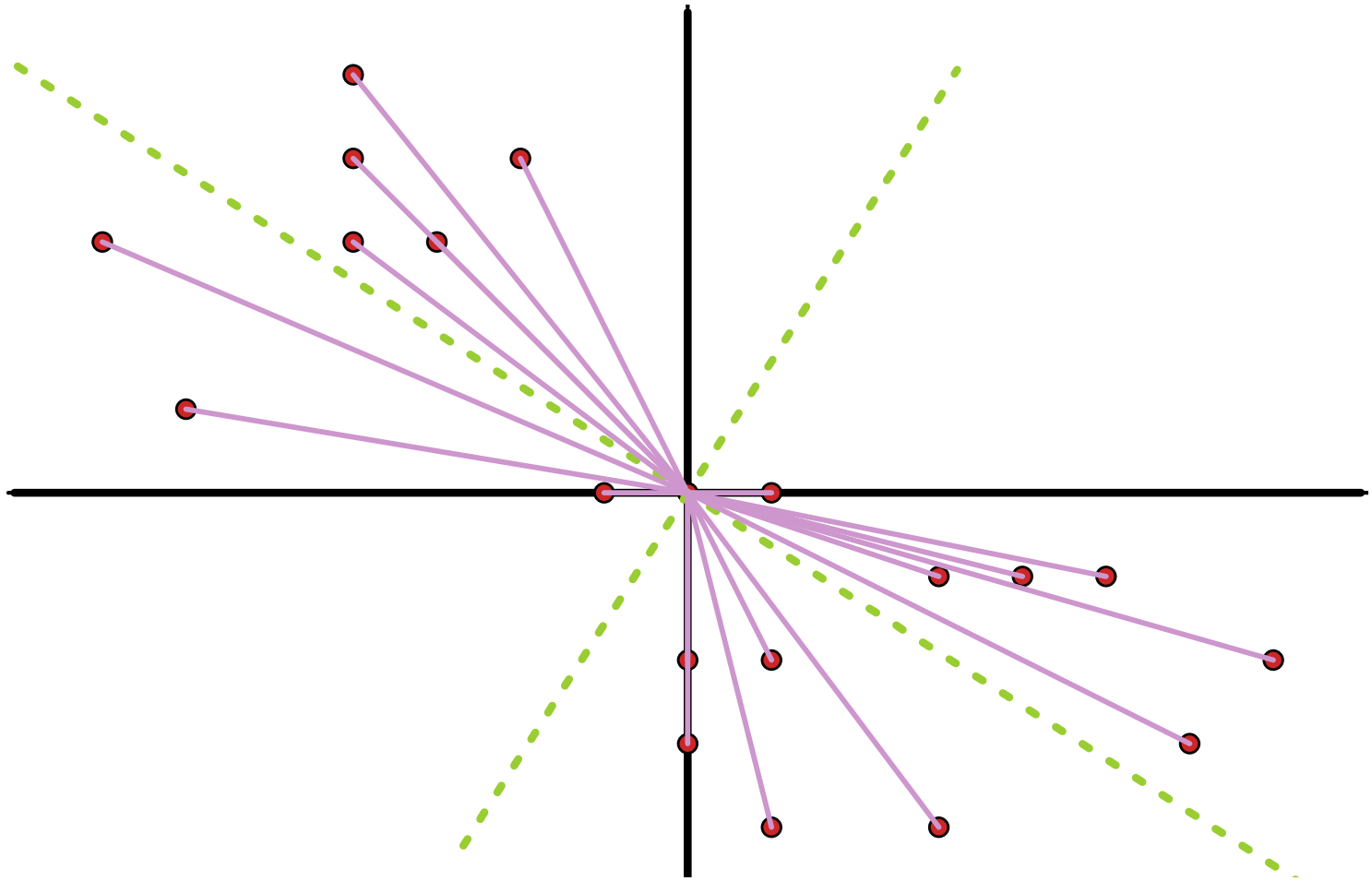

**Squared distance from "Listrac" to center of the space From variable**

$$3^2 + (-4)^2 = 9 + 16 = 25$$

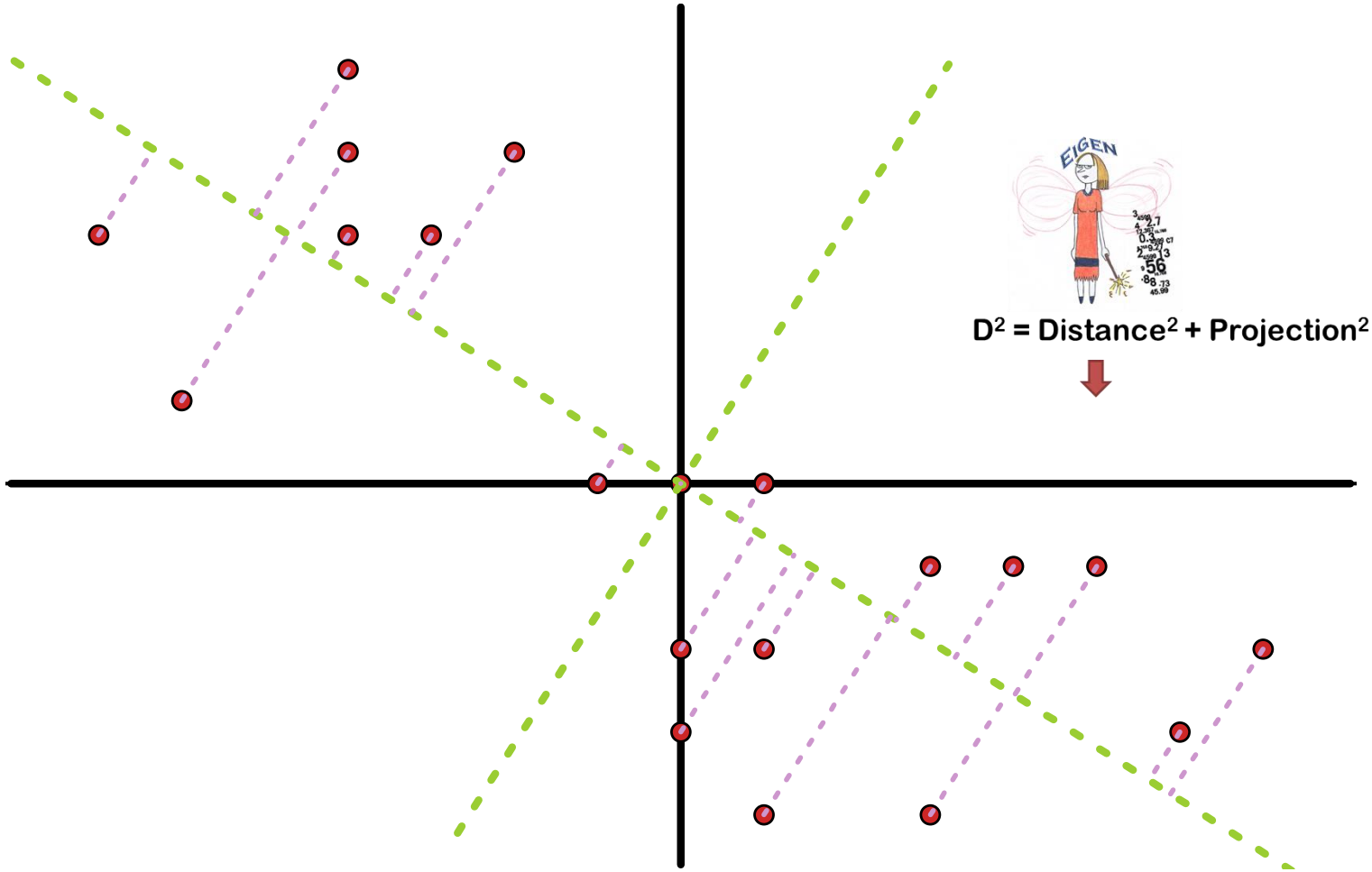


# All observations: total inertia

# Total Variance (aka Inertia)

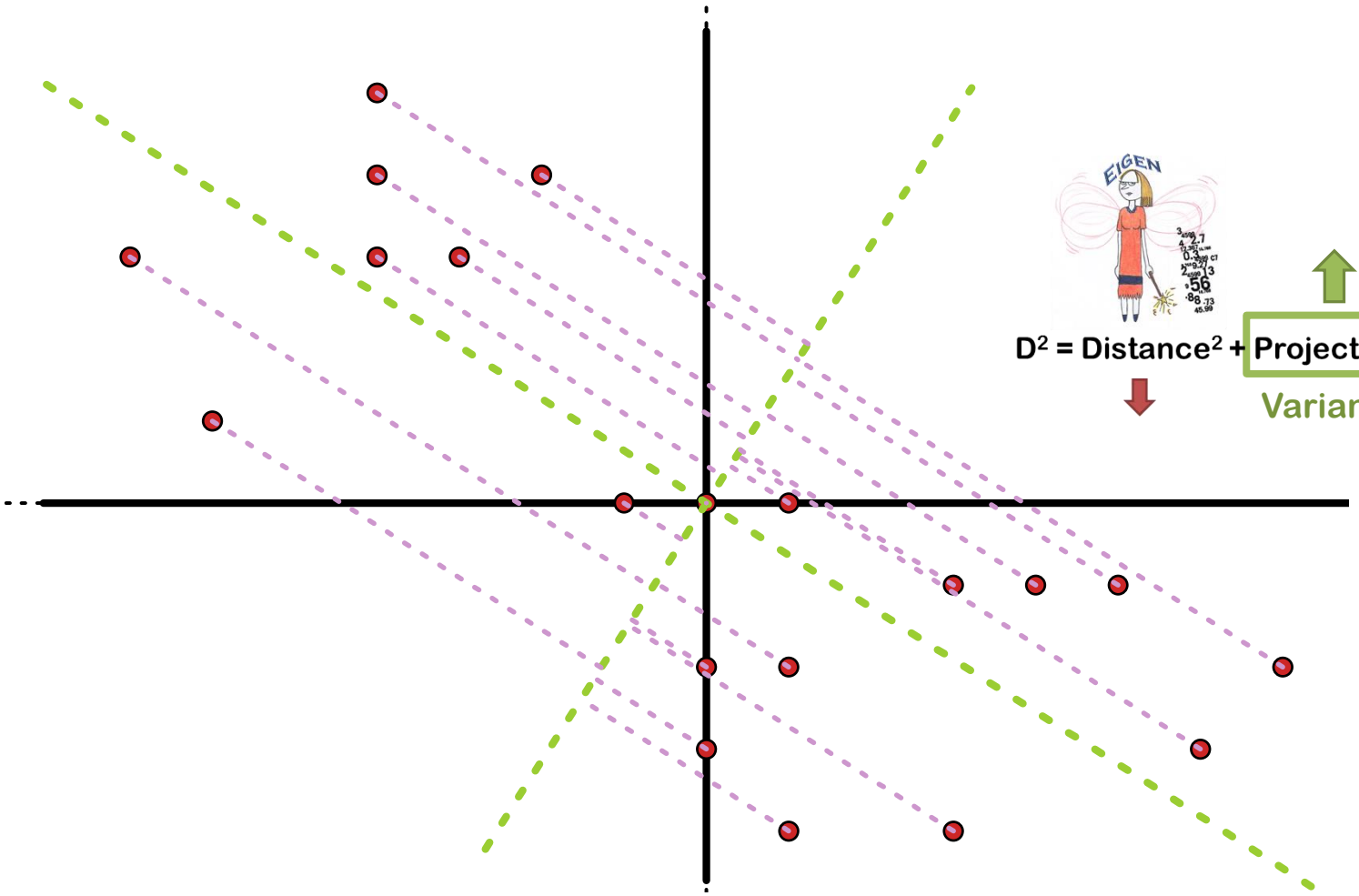


# Min Error Variance (aka Inertia)



# Pythagorean Magic

# Max Variance (aka Inertia)






$D^2 = \text{Distance}^2 + \text{Projection}^2$

Variance

$D^2 = A^2 + B^2$

# Pythagorean magic:

## Smallest sum of squared distances

## same as

## Largest squared projections (variance)

$D^2 = Distance^2 + Projection^2$

# Min Error = max Variance

Variance is called:

an eigen-value

A characteristic root

Latent root

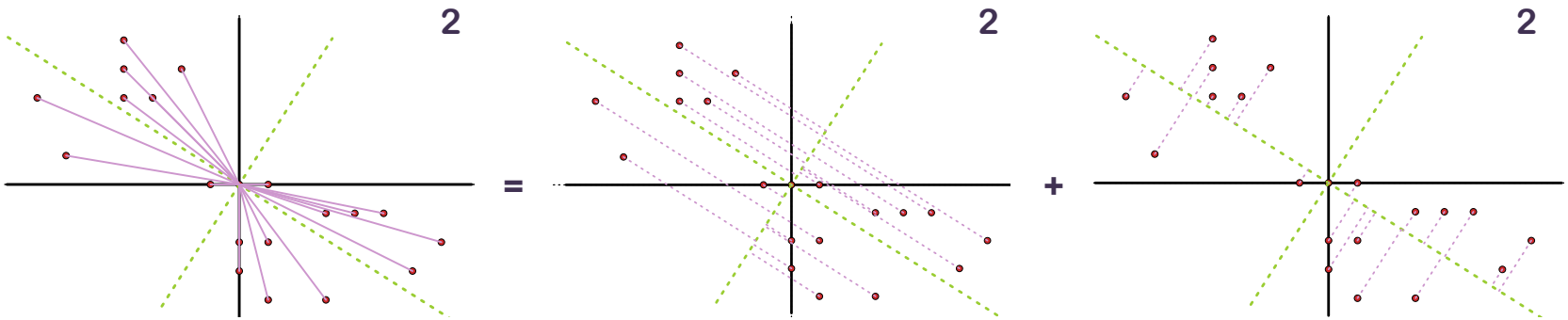
Proper root

French: “valeur propre”

$$D^2 = A^2 + B^2$$

# PYTHAGOREAN MAGIC:

## TOTAL INERTIA = SUM OF EIGENVALUES



## How Good?

# Good components explain a lot

| Component | $\lambda_j$ (eigenvalue) | Cumulated (eigenvalues) | Percent of inertia | Cumulated (percentage) |
|---|---|---|---|---|
| 1 | 392 | 392 | 83.29 | 83.29 |
| 2 | 52 | 444 | 11.71 | 100.00 |

# Compared to

| Variable | $SS_j$ | Cumulative $SS_j$ | Percent of inertia | Cumulative (percentage) |
|---|---|---|---|---|
| Astringent | 294 | 294 | 66.22 | 66.22 |
| Sugar | 150 | 444 | 33.78 | 100.00 |

# Eigen-fairy: You cannot better PCA!

# How to represent the variables?

## Part 1. Correlations aka Loadings

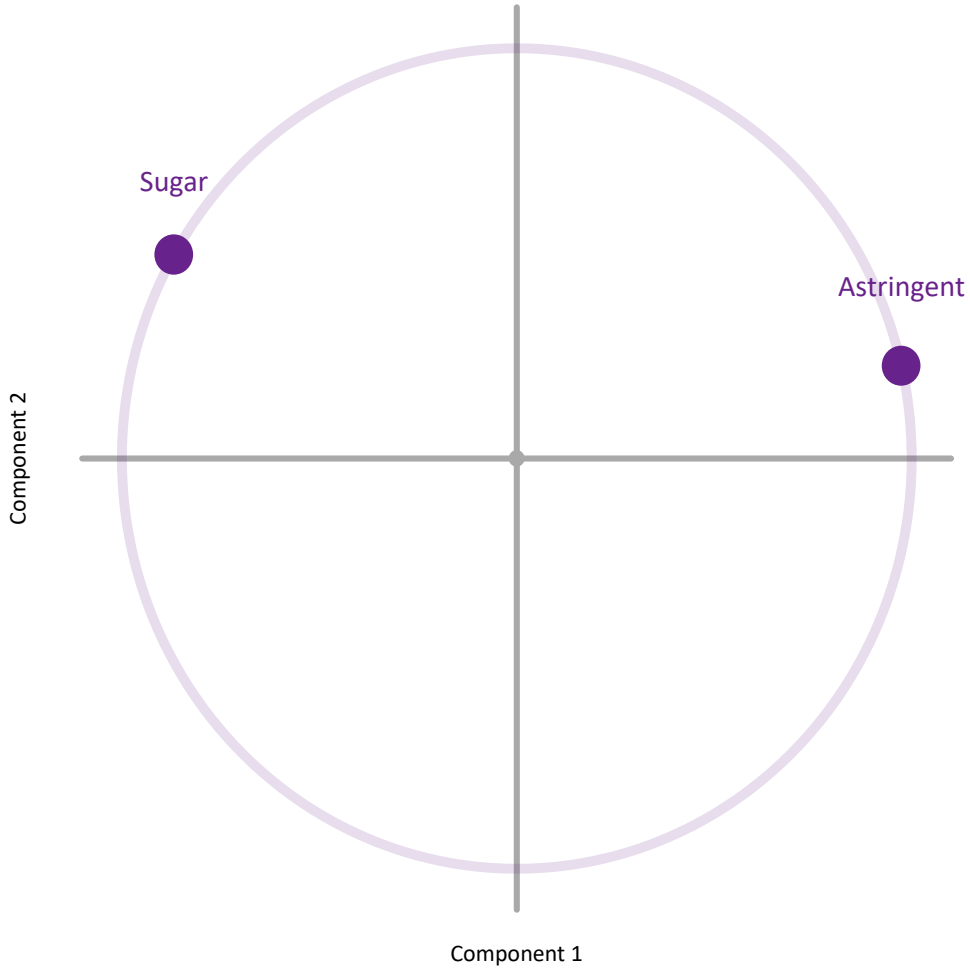
## Correlation between variables and factor scores (très French, by the way)

# First Compute the Correlations

*Y*: Astringent. *W*: Sugar

| | Loadings | |
|---|---|---|
| Component | *Y* | *W* |
| 1 | .9742 | –.8670 |
| 2 | .2258 | .4967 |
| Σ | | |

# Circle of Correlation



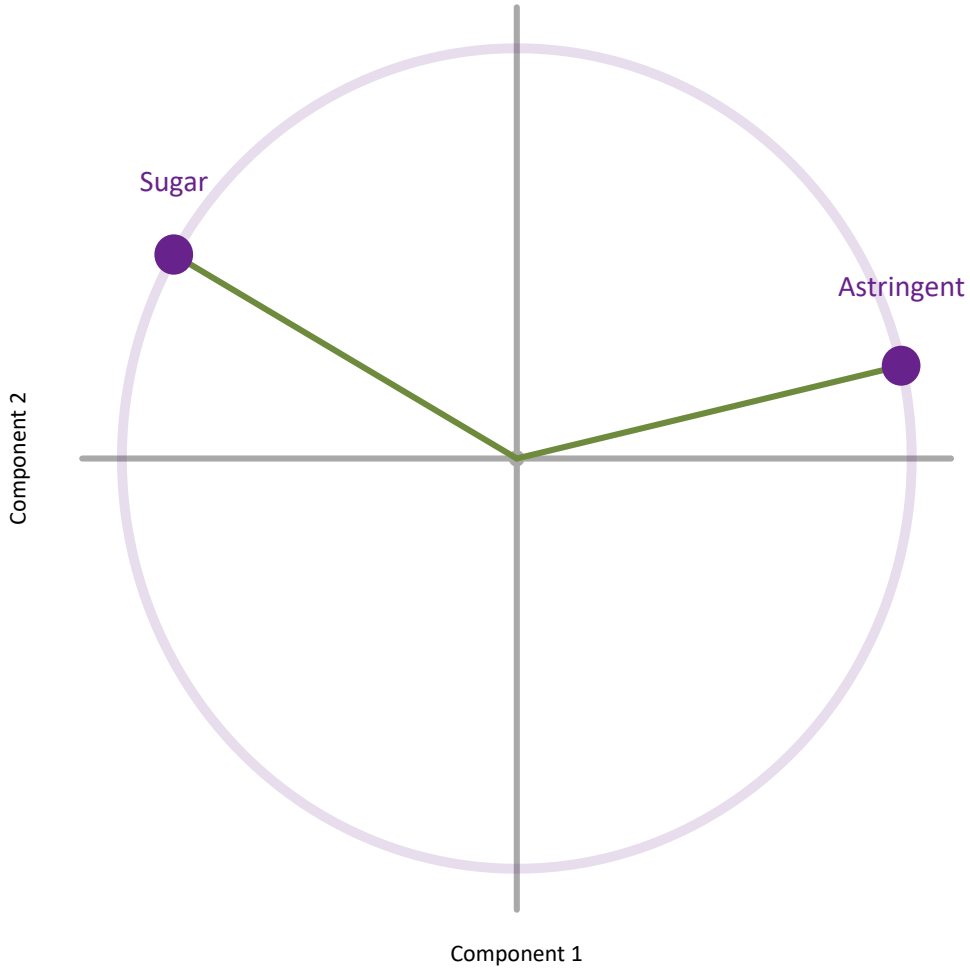
# Descartes's Circle

## Why a Circle of radius 1?

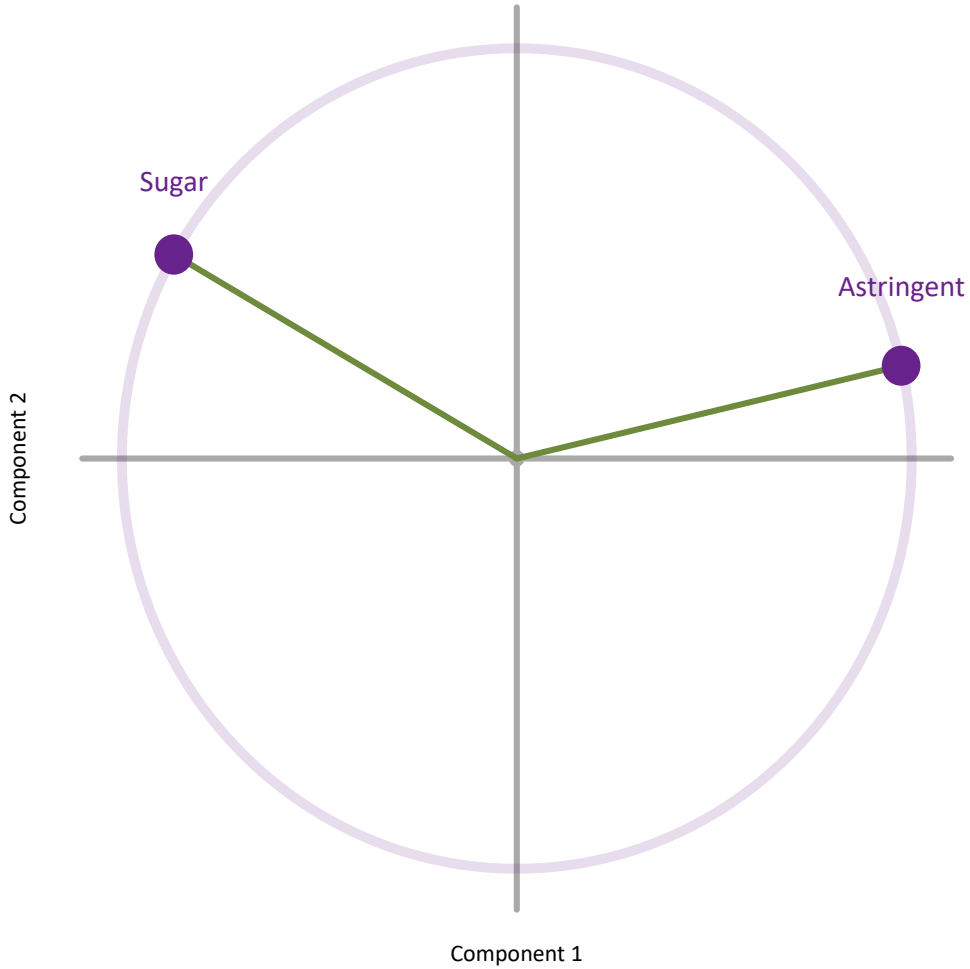
## This is Descartes's fault

## A circle is defined as $x^2 + y^2 = \text{radius}$

# Circle of Correlation

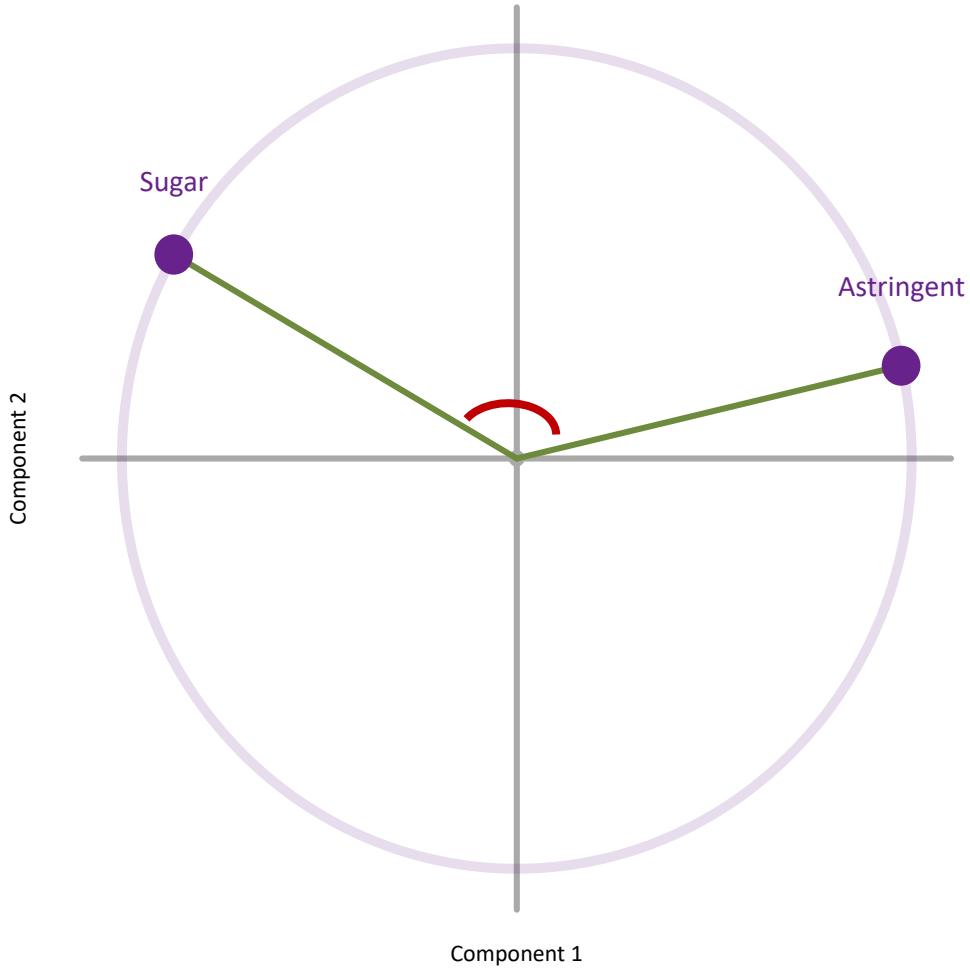


# Circle of Correlation



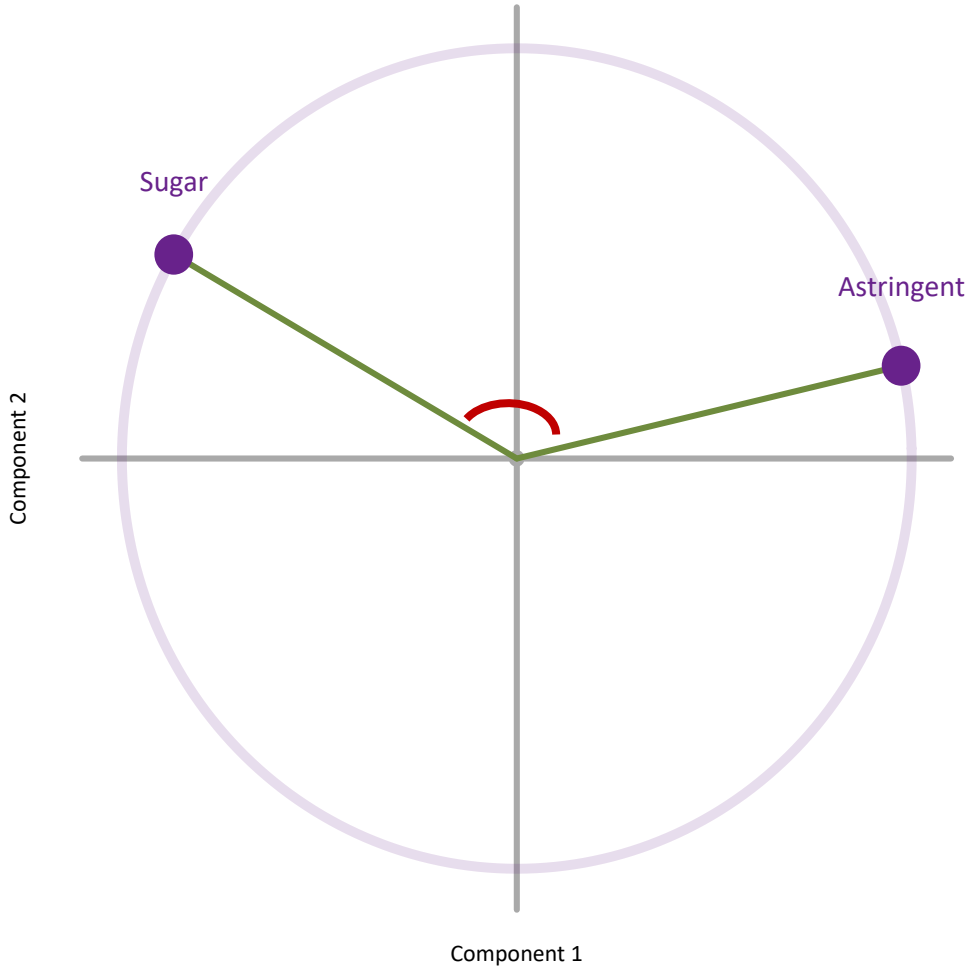

## How to interpret:

# COSINES APPROXIMATE CORRELATION




## HOW TO INTERPRET:

# COSINES APPROXIMATE CORRELATION

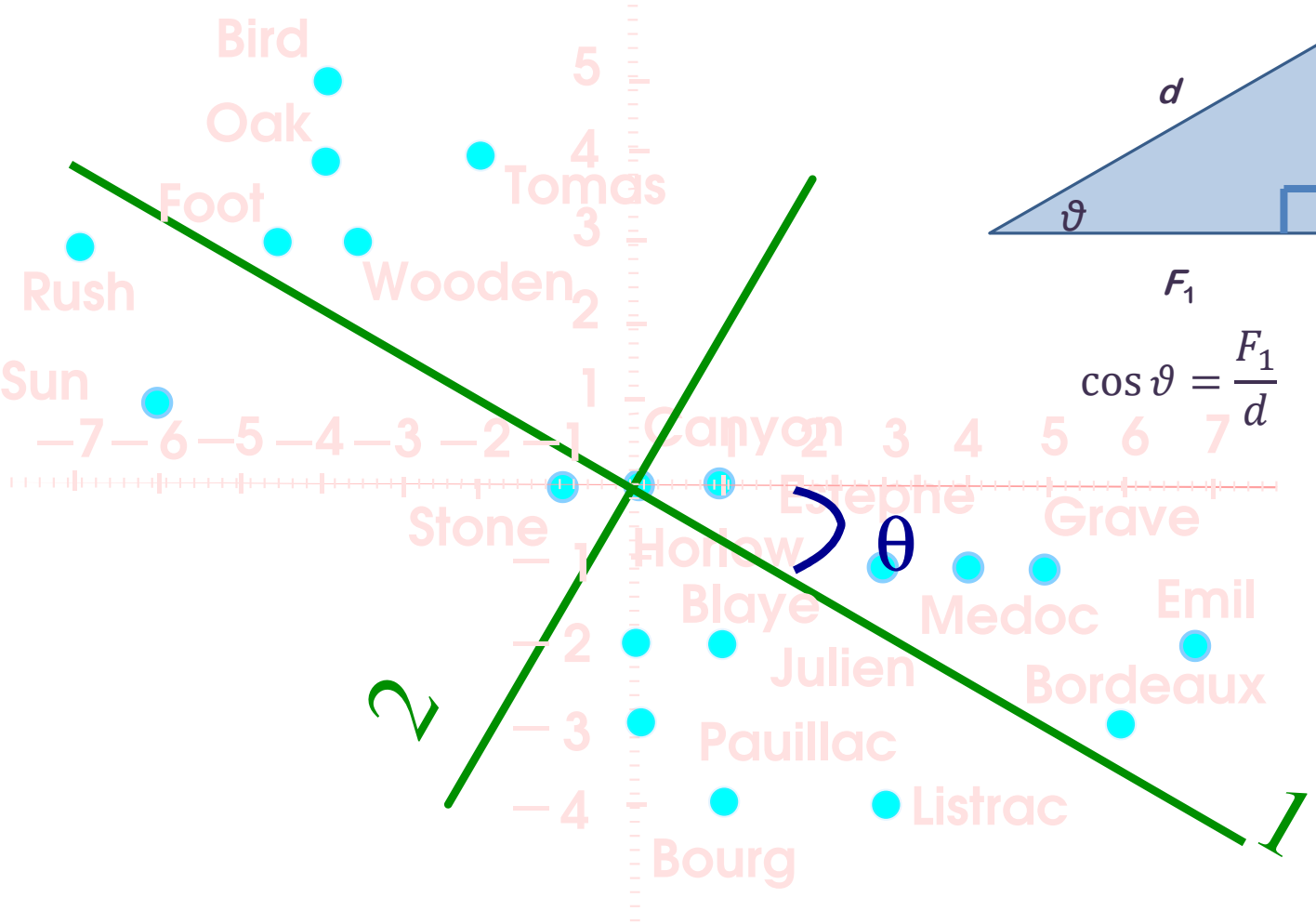



## HOW TO INTERPRET: ANGLE = 137°, $R = -.73$

# How to represent the variables?

## Part 2. Cosines (angles) as Loadings

# PART 2. COSINES (ANGLES) AS LOADINGS







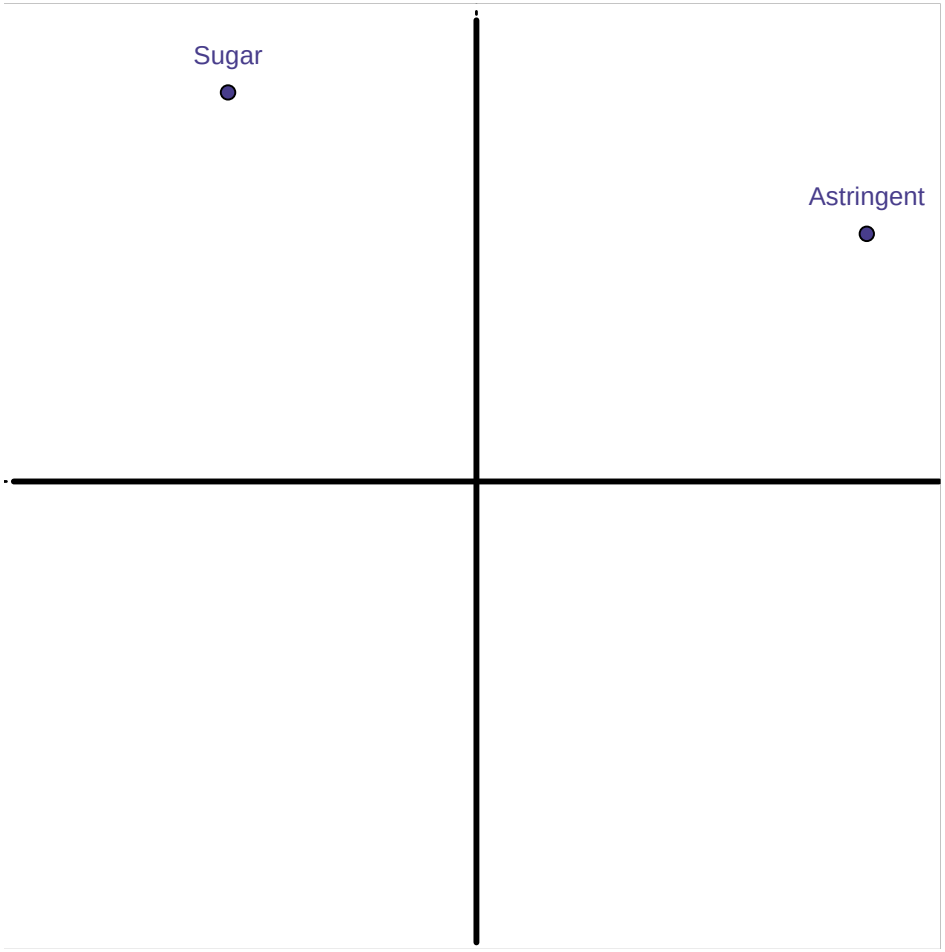
$$\cos\vartheta = \frac{F_1}{d}$$

# Loadings as cosines

*Y*: Astringent. *W*: Sugar

| Component | Loadings | | Squared Loadings | | Σ |
|---|---|---|---|---|---|
| | *Y* | *W* | *Y* | *W* | |
| 1 | .8437 | −.5369 | .7118 | .2882 | 1 |
| 2 | .5369 | .8437 | .2882 | .7118 | 1 |

# Plot them!



# Loadings as "slices of Inertia"

*Y*: Astringent. *W*: Sugar

*Sum of row: Component Inertia (eigenvalue)*

| | Loadings | | Squared Loadings | | Σ |
|---|---|---|---|---|---|
| Component | *Y* | *W* | *Y* | *W* | |
| 1 | 16.70 | –10.63 | 279 | 113 | **392** |
| 2 | 3.87 | 6.08 | 15 | 37 | **52** |
| | | | **294** | **150** | **444** |

*Sum of columns: Variable Inertia*

```
Exposition: resPCA$ExPosition.Data$fj
```

# What to do with a new wine? (Trius, a wine from Ontario)


Rated:
Sugar: 3/20
Astringent: 12/20

Centered Rating:
Sugar: 3 – 6 = –3
Astringent: 12 – 8 = 4

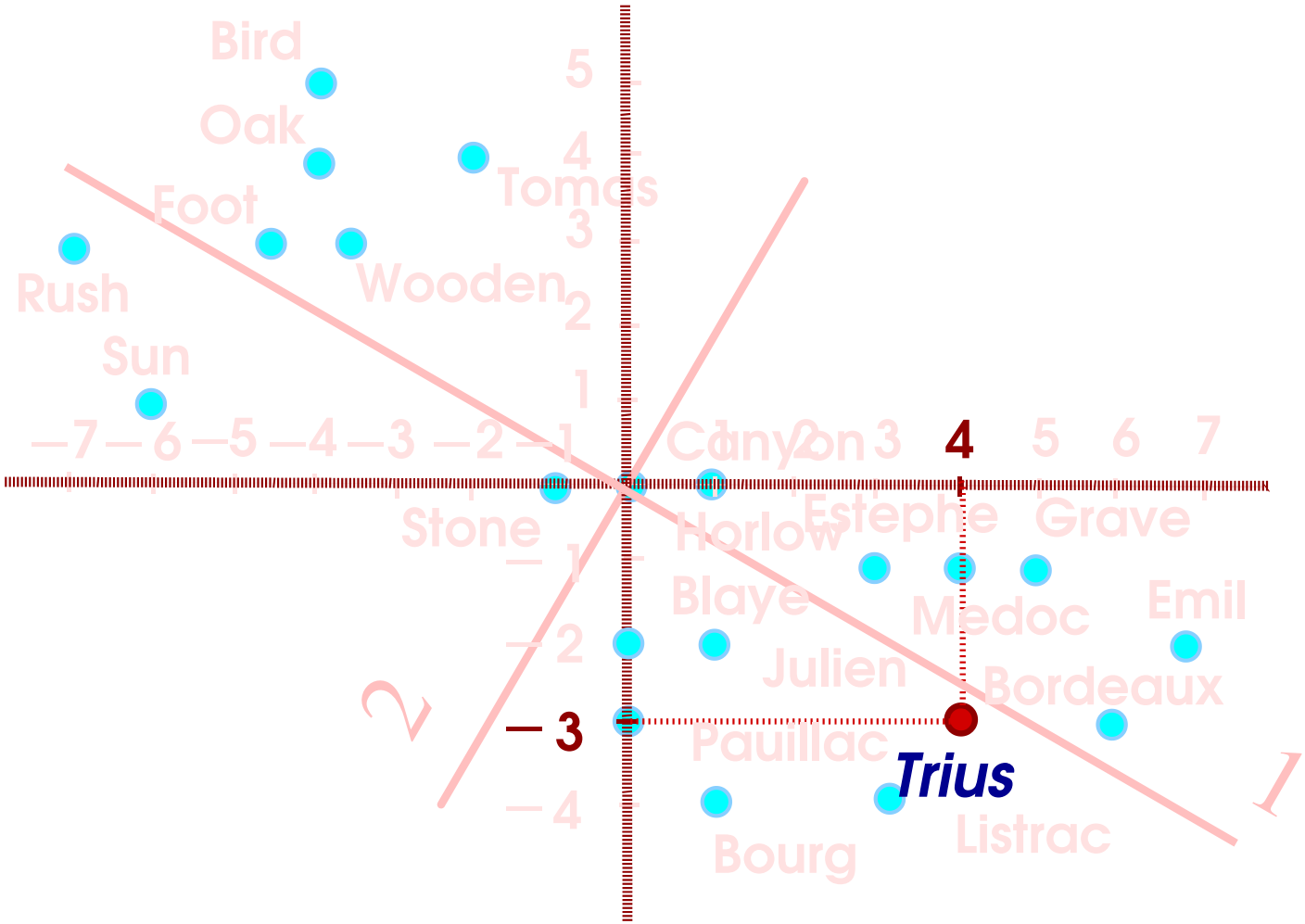
# This is called:

A supplementary observation

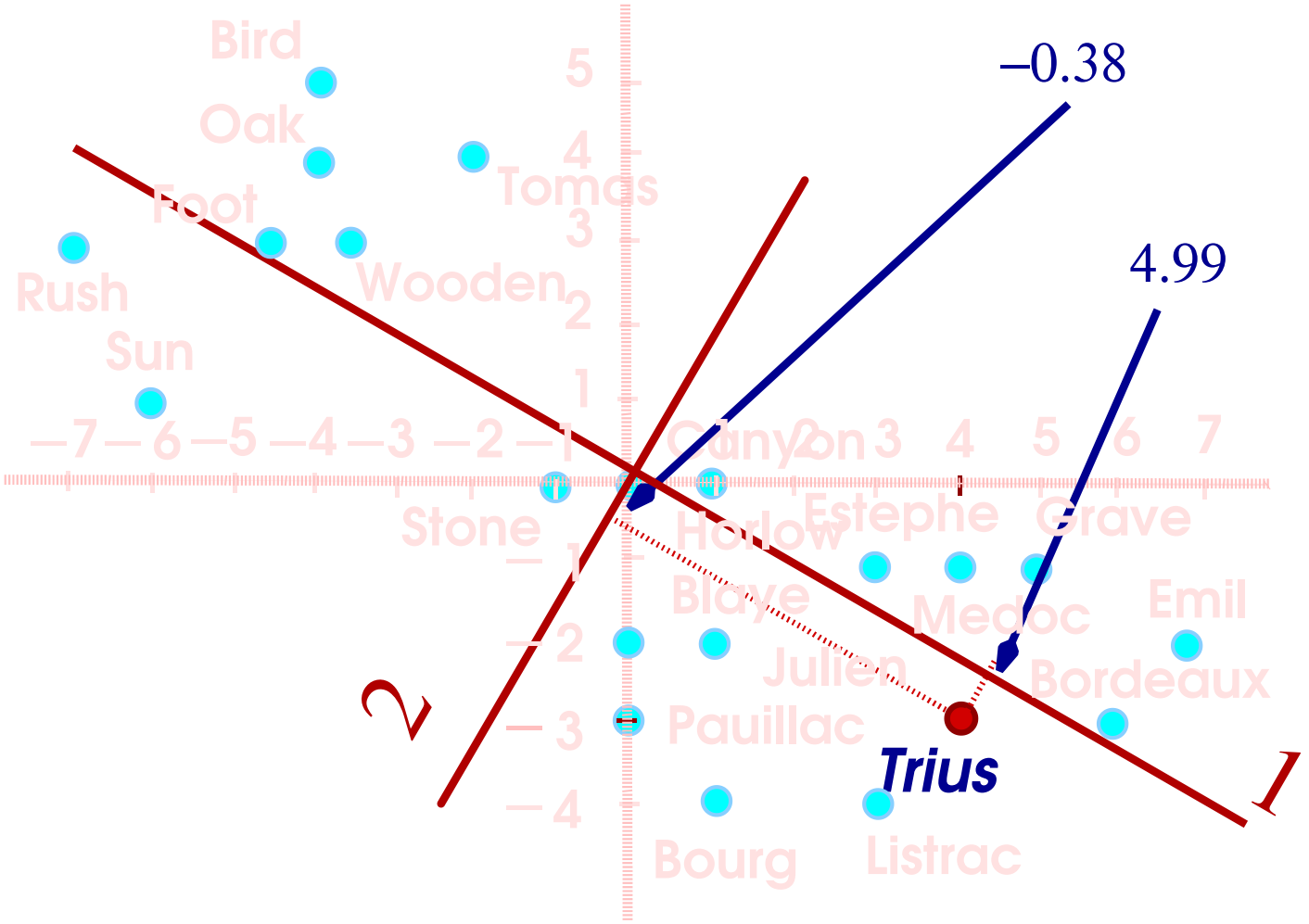
An Illustrative observation

An "out of sample" observation

# Just plot it in the space

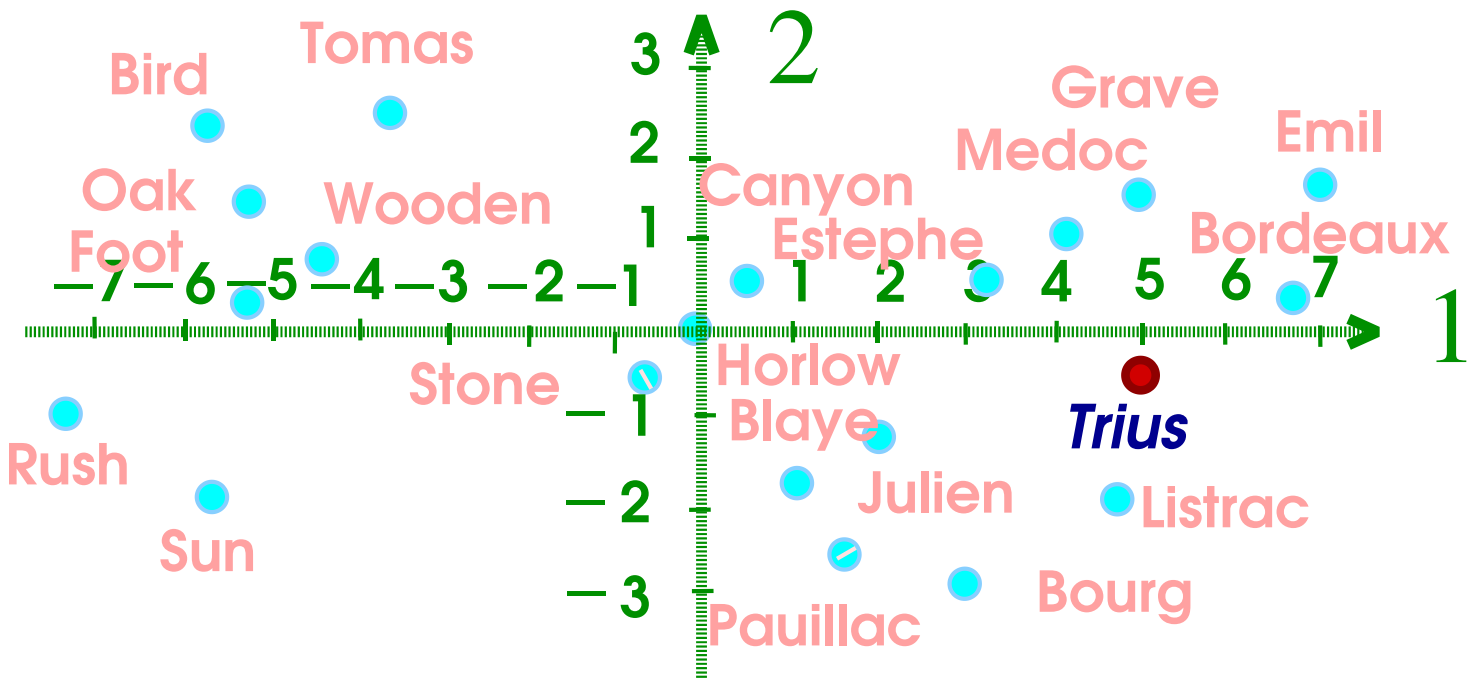


# Back-project on components



# Trius

## "Horizontalize" Component 1




## Looks like Trius is Cabernet + Merlot (and it is!)

# What about new variables?

# 20 Wines: All data

Here are the data

In case you had forgotten.

| (Long) Name | 1. Bordeaux | 2. Black Stone | 3. Listrac | 4. Canyon Creek | 5. Côtes de Bourg | 6. Foot Hill | 7. Horlow | 8. St. Estèphe | 9. Wooden Hill | 11. Côtes de Blaye | 11. Sun Set | 12. Black Bird | 13. Médoc | 14 St. Julien | 15. Pauillac | 16. Gold Rush | 17. Oak Ville | 18. Grave | 19. St. Emillion | 20. Tomasello |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Short Name | 1. Bordeaux | 2. Stone | 3. Listrac | 4. Canyon | 5. Bourg | 6. Hill | 7. Horlow | 8. Estephe | 9. Wooden | 10. Blaye | 11. Sun | 12. Bird | 13. Medoc | 14. Julien | 15. Pauillac | 16. Rush | 17. Oak | 18. Grave | 19. Emil | 20. Tomas |
| Y: Sugar | 3 | 6 | 2 | 6 | 2 | 9 | 6 | 5 | 9 | 4 | 7 | 11 | 5 | 4 | 3 | 9 | 10 | 5 | 4 | 10 |
| W: Astringency | 14 | 7 | 11 | 9 | 9 | 4 | 8 | 11 | 5 | 8 | 2 | 4 | 12 | 9 | 8 | 1 | 4 | 13 | 15 | 6 |
| Origin | F | U | F | U | F | U | U | F | U | F | U | U | F | F | F | U | U | F | F | U |
| Fruity | | | F | | F | | | | | F | F | | | F | F | F | | | | |
| Woody | | | | | | | | | | | | W | W | | | | | W | W | W |
| Acid | 2 | 5 | 1 | 1 | 9 | 1 | 2 | 2 | 1 | 1 | 2 | 1 | 2 | 9 | 14 | 2 | 1 | 2 | 1 | 1 |
| Bitter | 8 | 3 | 16 | 3 | 11 | 1 | 1 | 9 | 1 | 8 | 3 | 2 | 9 | 12 | 10 | 2 | 1 | 10 | 7 | 3 |

## What about new variables?

## This is called:

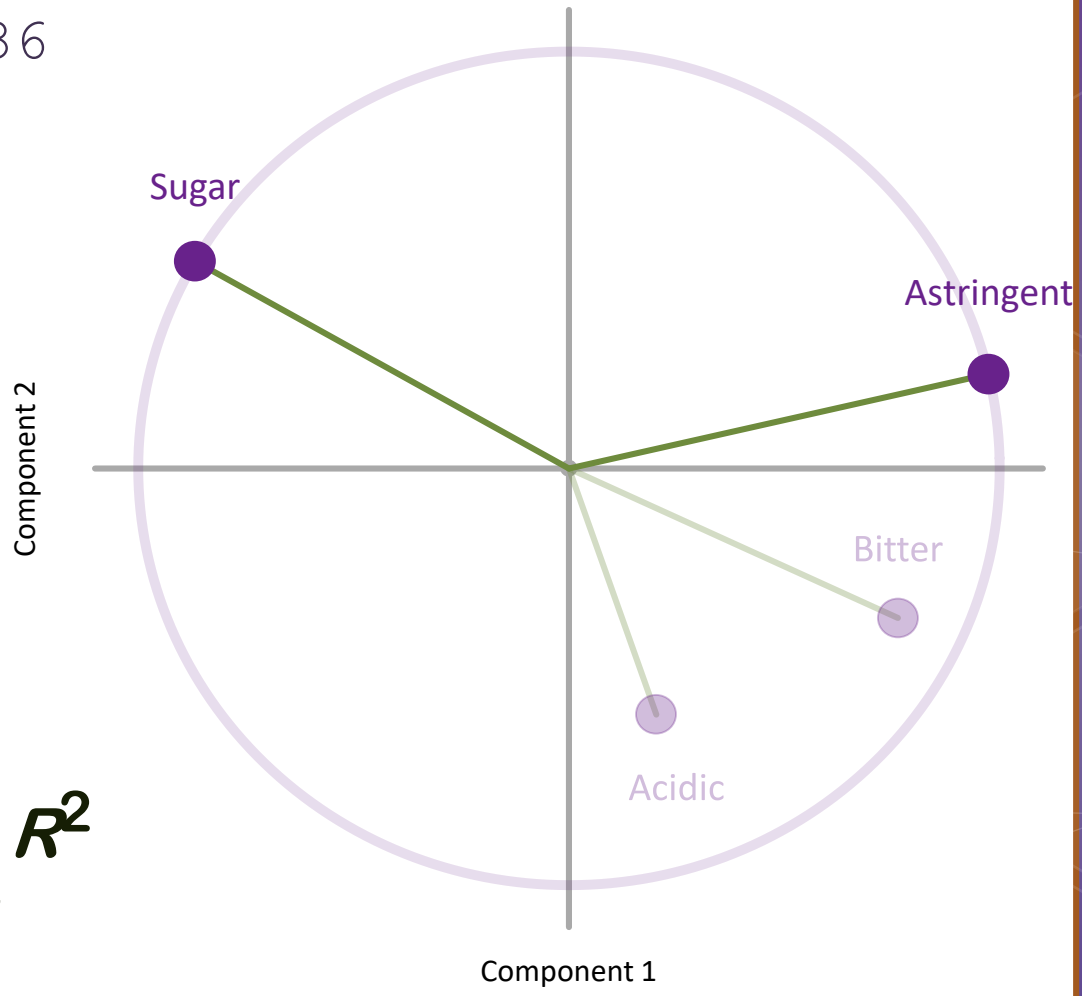
## A supplementary variable

## An Illustrative variable

## How: Compute Correlation → Loadings

# Plot them

| | PC1 | PC2 |
|---|---|---|
| *Acidic* | .20 | -.59 |
| *Bitter* | .76 | -.36 |




## Correlation:

***Supplementary $R^2$ do not sum to 1***

# Mid-talk wrap-up

Maps are Best!

Nearby observations are alike


Cosines are for variables: [–1, 1]

Factor scores are best mixture

Good factors explain a lot

"Merci" to the eigen-fairy

# What are the important Components

# GOOD COMPONENTS EXPLAIN A LOT

| Component | $\lambda_i$ (eigenvalue) | Cumulated (eigenvalues) | Percent of inertia | Cumulated (percentage) |
|---|---|---|---|---|
| 1 | 392 | 392 | 83.29 | 83.29 |
| 2 | 52 | 444 | 11.71 | 100.00 |

# COMPARED TO

| Variable | $SS_i$ | Cumulative $SS_i$ | Percent of inertia | Cumulative (percentage) |
|---|---|---|---|---|
| Astringent | 294 | 294 | 66.22 | 66.22 |
| Sugar | 150 | 444 | 33.78 | 100.00 |

# EIGEN-FAIRY: YOU CANNOT BETTER PCA!

# What are the important Components

## Rule of thumb:

## A Good component explains more than the average Inertia.

# Important Observations

## First a refresher

# Eigenvalues

## Sum of squared projections

| | $Y$ | $W$ | $y$ | $w$ | $F_1$ | $F_2$ | $F_1^2$ | $F_2^2$ |
|---|---|---|---|---|---|---|---|---|
| Bordeaux | 3 | 14 | −3 | 6 | 6.67 | 0.69 | 44.52 | 0.48 |
| Black Stone | 6 | 7 | 0 | −1 | −0.84 | −0.54 | 0.71 | 0.29 |
| Listrac | 2 | 11 | −4 | 3 | 4.68 | −1.76 | 21.89 | 3.11 |
| Canyon Creek | 6 | 9 | 0 | 1 | 0.84 | 0.54 | 0.71 | 0.29 |
| Côtes de Bourg | 2 | 9 | −4 | 1 | 2.99 | −2.84 | 8.95 | 8.05 |
| Foot Hill | 9 | 4 | 3 | −4 | −4.99 | 0.38 | 24.85 | 0.15 |
| Horlow | 6 | 8 | 0 | 0 | 0.00 | 0.00 | 0 | 0.00 |
| St. Estphe | 5 | 11 | −1 | 3 | 3.07 | 0.77 | 9.41 | 0.59 |
| Wooden Hill | 9 | 5 | 3 | −3 | −4.14 | 0.92 | 17.15 | 0.85 |
| Côtes de Blaye | 4 | 8 | −2 | 0 | 1.07 | −1.69 | 1.15 | 2.85 |
| Sun Set | 7 | 2 | 1 | −6 | −5.60 | −2.38 | 31.35 | 5.65 |
| Black Bird | 11 | 4 | 5 | −4 | −6.06 | 2.07 | 36.71 | 4.29 |
| Médoc | 5 | 12 | −1 | 4 | 3.91 | 1.30 | 15.30 | 1.70 |
| St Julien | 4 | 9 | −2 | 1 | 1.92 | −1.15 | 3.68 | 1.32 |
| Pauillac | 3 | 8 | −3 | 0 | 1.61 | −2.53 | 2.59 | 6.41 |
| Gold Rush | 9 | 1 | 3 | −7 | −7.52 | −1.23 | 56.49 | 1.51 |
| Oak Ville | 10 | 4 | 4 | −4 | −5.52 | 1.23 | 30.49 | 1.51 |
| Grave | 5 | 13 | −1 | 5 | 4.76 | 1.84 | 22.61 | 3.39 |
| St Emilion | 4 | 15 | −2 | 7 | 6.98 | 2.07 | 48.71 | 4.29 |
| Tomasello | 10 | 6 | 4 | −2 | −3.83 | 2.30 | 14.71 | 5.29 |
| $\Sigma$ | 120 | 160 | 0 | 0 | 0 | 0 | 392 $\lambda_1$ | 52 $\lambda_2$ |

## Important Observations:

Rule of thumb:

A Good observation “explains” more than the average Inertia.

How to find the “explained inertia?”

# First make it a proportion

| $F_1^2$ | $F_2^2$ |
|---|---|
| 44.52 | 0.48 |
| 0.71 | 0.29 |
| 21.89 | 3.11 |
| 0.71 | 0.29 |
| 8.95 | 8.05 |
| 24.85 | 0.15 |
| 0.00 | 0.00 |
| 9.41 | 0.59 |
| 17.15 | 0.85 |
| 1.15 | 2.85 |
| 31.35 | 5.65 |
| 36.71 | 4.29 |
| 15.30 | 1.70 |
| 3.68 | 1.32 |
| 2.59 | 6.41 |
| 56.49 | 1.51 |
| 30.49 | 1.51 |
| 22.61 | 3.39 |
| 48.71 | 4.29 |
| 14.71 | 5.29 |
| $\lambda$ 392 | 52 |

44.52 / 392 = .114

## Call this ratio: the contribution

## Important observations:

## contribution >

# Factor scores

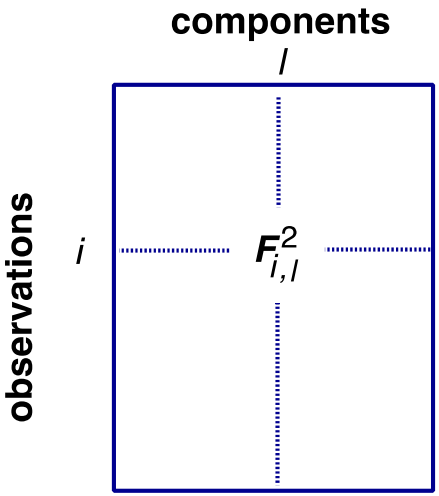
## Eigenvalues

## Sum of squared projections

(1/20) = .05

| | Y | W | y | w | $F_1$ | $F_2$ | ctr$_1$ × 100 | ctr$_2$ × 100 | $F_1^2$ | $F_2^2$ | $d^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bordeaux | 3 | 14 | −3 | 6 | 6.67 | 0.69 | 11 | 1 | 44.52 | 0.48 | 45 |
| Black Stone | 6 | 7 | 0 | −1 | −0.84 | −0.54 | 0 | 1 | 0.71 | 0.29 | 1 |
| Listrac | 2 | 11 | −4 | 3 | 4.68 | −1.76 | 6 | 6 | 21.89 | 3.11 | 25 |
| Canyon Creek | 6 | 9 | 0 | 1 | 0.84 | 0.54 | 0 | 1 | 0.71 | 0.29 | 1 |
| Côtes de Bourg | 2 | 9 | −4 | 1 | 2.99 | −2.84 | 2 | 15 | 8.95 | 8.05 | 17 |
| Foot Hill | 9 | 4 | 3 | −4 | −4.99 | 0.38 | 6 | 0 | 24.85 | 0.15 | 25 |
| Horlow | 6 | 8 | 0 | 0 | 0.00 | 0.00 | 0 | 0 | 0 | 0.00 | 0 |
| St. Estphe | 5 | 11 | −1 | 3 | 3.07 | 0.77 | 3 | 1 | 9.41 | 0.59 | 10 |
| Wooden Hill | 9 | 5 | 3 | −3 | −4.14 | 0.92 | 5 | 2 | 17.15 | 0.85 | 18 |
| Côtes de Blaye | 4 | 8 | −2 | 0 | 1.07 | −1.69 | 0 | 5 | 1.15 | 2.85 | 4 |
| Sun Set | 7 | 2 | 1 | −6 | −5.60 | −2.38 | 8 | 11 | 31.35 | 5.65 | 37 |
| Black Bird | 11 | 4 | 5 | −4 | −6.06 | 2.07 | 9 | 8 | 36.71 | 4.29 | 41 |
| Médoc | 5 | 12 | −1 | 4 | 3.91 | 1.30 | 4 | 3 | 15.30 | 1.70 | 17 |
| St Julien | 4 | 9 | −2 | 1 | 1.92 | −1.15 | 1 | 3 | 3.68 | 1.32 | 5 |
| Pauillac | 3 | 8 | −3 | 0 | 1.61 | −2.53 | 1 | 12 | 2.59 | 6.41 | 9 |
| Gold Rush | 9 | 1 | 3 | −7 | −7.52 | −1.23 | 14 | 3 | 56.49 | 1.51 | 58 |
| Oak Ville | 10 | 4 | 4 | −4 | −5.52 | 1.23 | 8 | 3 | 30.49 | 1.51 | 32 |
| Grave | 5 | 13 | −1 | 5 | 4.76 | 1.84 | 6 | 7 | 22.61 | 3.39 | 26 |
| St Emilion | 4 | 15 | −2 | 7 | 6.98 | 2.07 | 12 | 8 | 48.71 | 4.29 | 53 |
| Tomasello | 10 | 6 | 4 | −2 | −3.83 | 2.30 | 4 | 10 | 14.71 | 5.29 | 20 |
| $\Sigma$ | 120 | 160 | 0 | 0 | 0 | 0 | 100 | 100 | 392 $\lambda_1$ | 52 $\lambda_2$ | 444 $\mathcal{G}$ |

# CONTRIBUTIONS IN A NUTSHELL



# Contributions: important observations for a component

What are the important components for an observation?

Good components have large cosines$^2$

# Factor scores

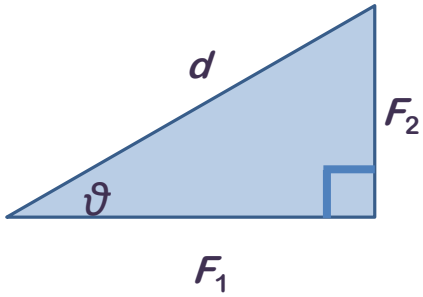
## Distance$^2$

## Sum of squared projections ($F_1^2 + F_2^2 = D^2$)

$$\frac{44.52}{45} = .99$$

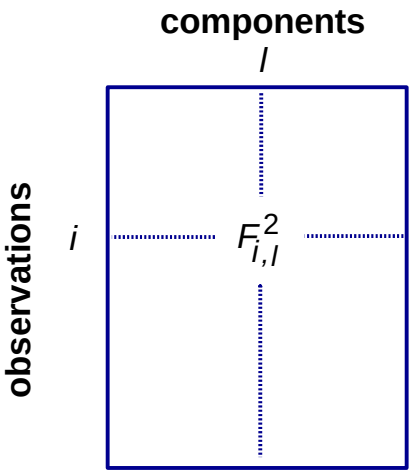
This is a cosine$^2$

$$\cos^2\vartheta = \frac{F_1^2}{d^2}$$




| | Y | W | y | w | $F_1$ | $F_2$ | $F_1^2$ | $F_2^2$ | $d^2$ | $\cos^2_1 \times 100$ | $\cos^2_2 \times 100$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Bordeaux | 3 | 14 | −3 | 6 | 6.67 | 0.69 | 44.52 | 0.48 | 45 | 99 | 1 |
| Black Stone | 6 | 7 | 0 | −1 | −0.84 | −0.54 | 0.71 | 0.29 | 1 | 71 | 29 |
| Listrac | 2 | 11 | −4 | 3 | 4.68 | −1.76 | 21.89 | 3.11 | 25 | 88 | 12 |
| Canyon Creek | 6 | 9 | 0 | 1 | 0.84 | 0.54 | 0.71 | 0.29 | 1 | 71 | 29 |
| Côtes de Bourg | 2 | 9 | −4 | 1 | 2.99 | −2.84 | 8.95 | 8.05 | 17 | 53 | 47 |
| Foot Hill | 9 | 4 | 3 | −4 | −4.99 | 0.38 | 24.85 | 0.15 | 25 | 99 | 1 |
| Horlow | 6 | 8 | 0 | 0 | 0.00 | 0.00 | 0 | 0.00 | 0 | 0 | 0 |
| St. Estphe | 5 | 11 | −1 | 3 | 3.07 | 0.77 | 9.41 | 0.59 | 10 | 94 | 6 |
| Wooden Hill | 9 | 5 | 3 | −3 | −4.14 | 0.92 | 17.15 | 0.85 | 18 | 95 | 5 |
| Côtes de Blaye | 4 | 8 | −2 | 0 | 1.07 | −1.69 | 1.15 | 2.85 | 4 | 29 | 71 |
| Sun Set | 7 | 2 | 1 | −6 | −5.60 | −2.38 | 31.35 | 5.65 | 37 | 85 | 15 |
| Black Bird | 11 | 4 | 5 | −4 | −6.06 | 2.07 | 36.71 | 4.29 | 41 | 90 | 10 |
| Médoc | 5 | 12 | −1 | 4 | 3.91 | 1.30 | 15.30 | 1.70 | 17 | 90 | 10 |
| St Julien | 4 | 9 | −2 | 1 | 1.92 | −1.15 | 3.68 | 1.32 | 5 | 74 | 26 |
| Pauillac | 3 | 8 | −3 | 0 | 1.61 | −2.53 | 2.59 | 6.41 | 9 | 29 | 71 |
| Gold Rush | 9 | 1 | 3 | −7 | −7.52 | −1.23 | 56.49 | 1.51 | 58 | 97 | 3 |
| Oak Ville | 10 | 4 | 4 | −4 | −5.52 | 1.23 | 30.49 | 1.51 | 32 | 95 | 5 |
| Grave | 5 | 13 | −1 | 5 | 4.76 | 1.84 | 22.61 | 3.39 | 26 | 87 | 13 |
| St Emilion | 4 | 15 | −2 | 7 | 6.98 | 2.07 | 48.71 | 4.29 | 53 | 92 | 8 |
| Tomasello | 10 | 6 | 4 | −2 | −3.83 | 2.30 | 14.71 | 5.29 | 20 | 74 | 26 |
| Σ | 120 | 160 | 0 | 0 | 0 | 0 | 392 | 52 | 444 | | |
| | | | | | | | $\lambda_1$ | $\lambda_2$ | $\mathcal{N}$ | | |

# COSINES$^2$ IN A NUTSHELL



## Variants of PCA

# And now for something completely

# not differently alike …

# The Unit Problem: Normalization.

# Choux et carottes

# Apples and oranges

# How?

# When *not* to normalize

# Ax Example: PCA with Covariance

**The story**

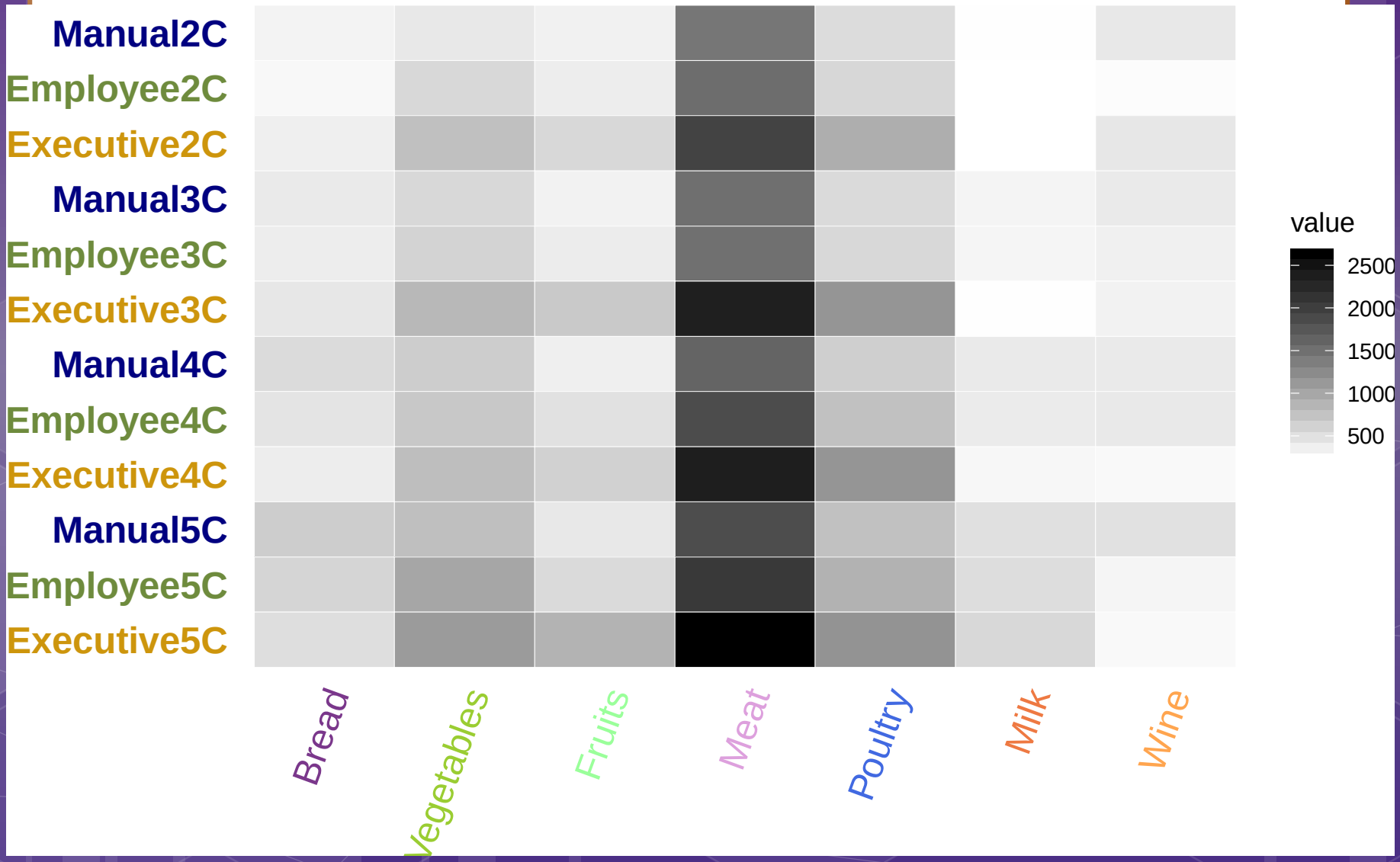
These data give the consumption in Francs of different "types of food" according to social class and number of children. The observations correspond to the average amount of money spent per month on a given type of food for a given social class and a given number of children. Because a franc spend on one item has the same value as one franc spent on another one, we want to keep the same unit of measurement for the complete space. Therefore we will analyze only the centered data, equivalently, we will analyze the covariance matrix.

# The data

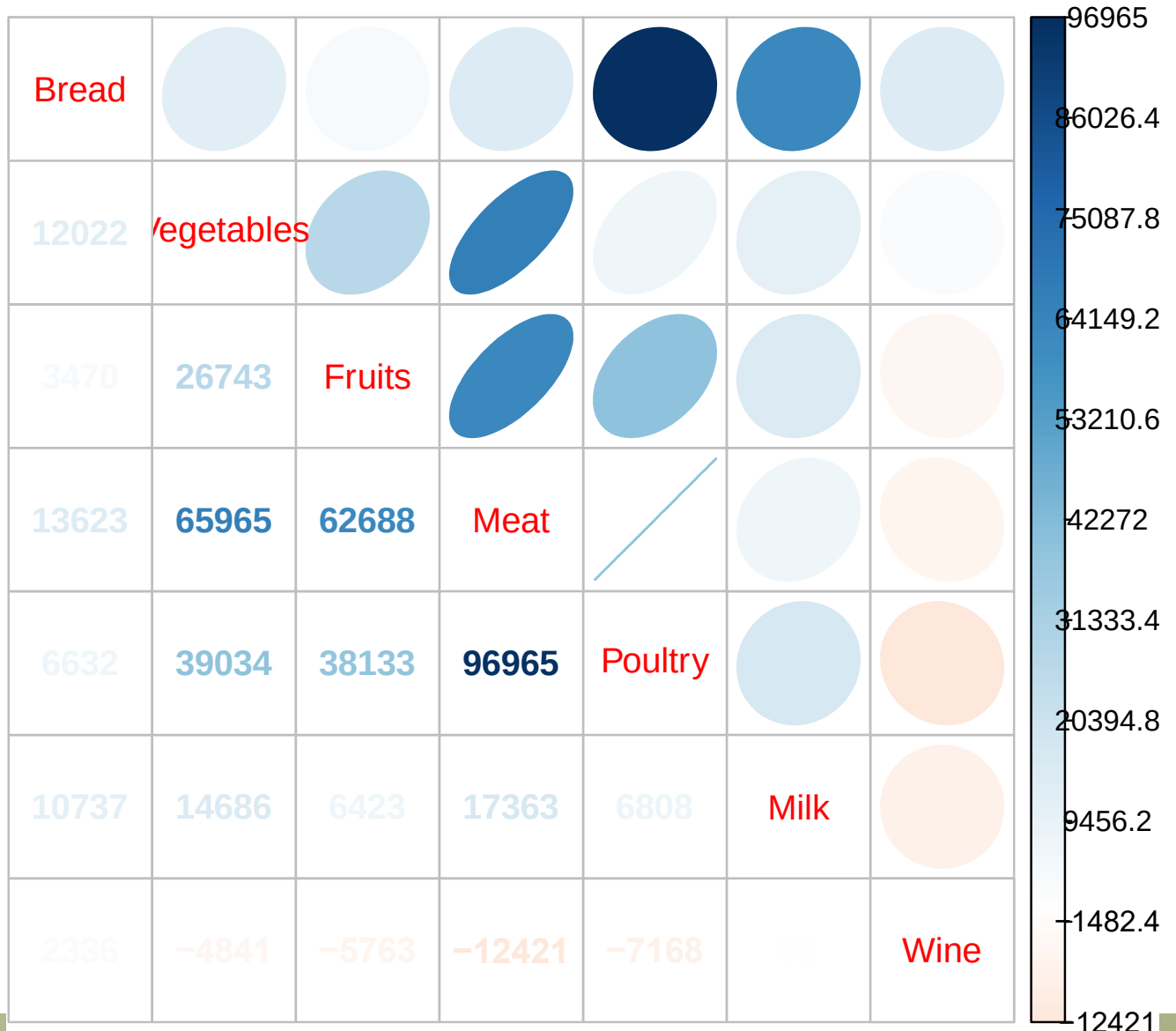
| Category | Bread | Vegies | Fruits | Meat | Pltr | Milk | Vine |
|---|---|---|---|---|---|---|---|
| Blue Clr, 2C. | 332 | 428 | 354 | 1437 | 526 | 247 | 427 |
| White Clr, 2C. | 293 | 559 | 388 | 1527 | 567 | 239 | 258 |
| Upper Cls, 2C. | 372 | 767 | 562 | 1948 | 927 | 235 | 433 |
| Blue Clr, 3C. | 406 | 563 | 341 | 1507 | 544 | 324 | 407 |
| White Clr, 3C. | 386 | 608 | 396 | 1501 | 558 | 319 | 363 |
| Upper Cls,3C. | 438 | 843 | 689 | 2345 | 1148 | 243 | 341 |
| Blue Clr, 4C. | 534 | 660 | 367 | 1620 | 638 | 414 | 407 |
| White Clr, 4C. | 460 | 699 | 484 | 1856 | 762 | 400 | 416 |
| Upper Cls, 4C. | 385 | 789 | 621 | 2366 | 1149 | 304 | 282 |
| Blue Clr, 5C. | 655 | 776 | 423 | 1848 | 759 | 495 | 486 |
| White Clr, 5C. | 584 | 995 | 548 | 2056 | 893 | 518 | 319 |
| Upper Cls, 5C. | 515 | 1097 | 887 | 2630 | 1167 | 561 | 284 |
| Mean | 447 | 732 | 505 | 1887 | 803 | 358 | 369 |
| $\hat{S}$ | 107 | 189 | 165 | 396 | 250 | 117 | 72 |

Table 13: The Data

# Data as picture



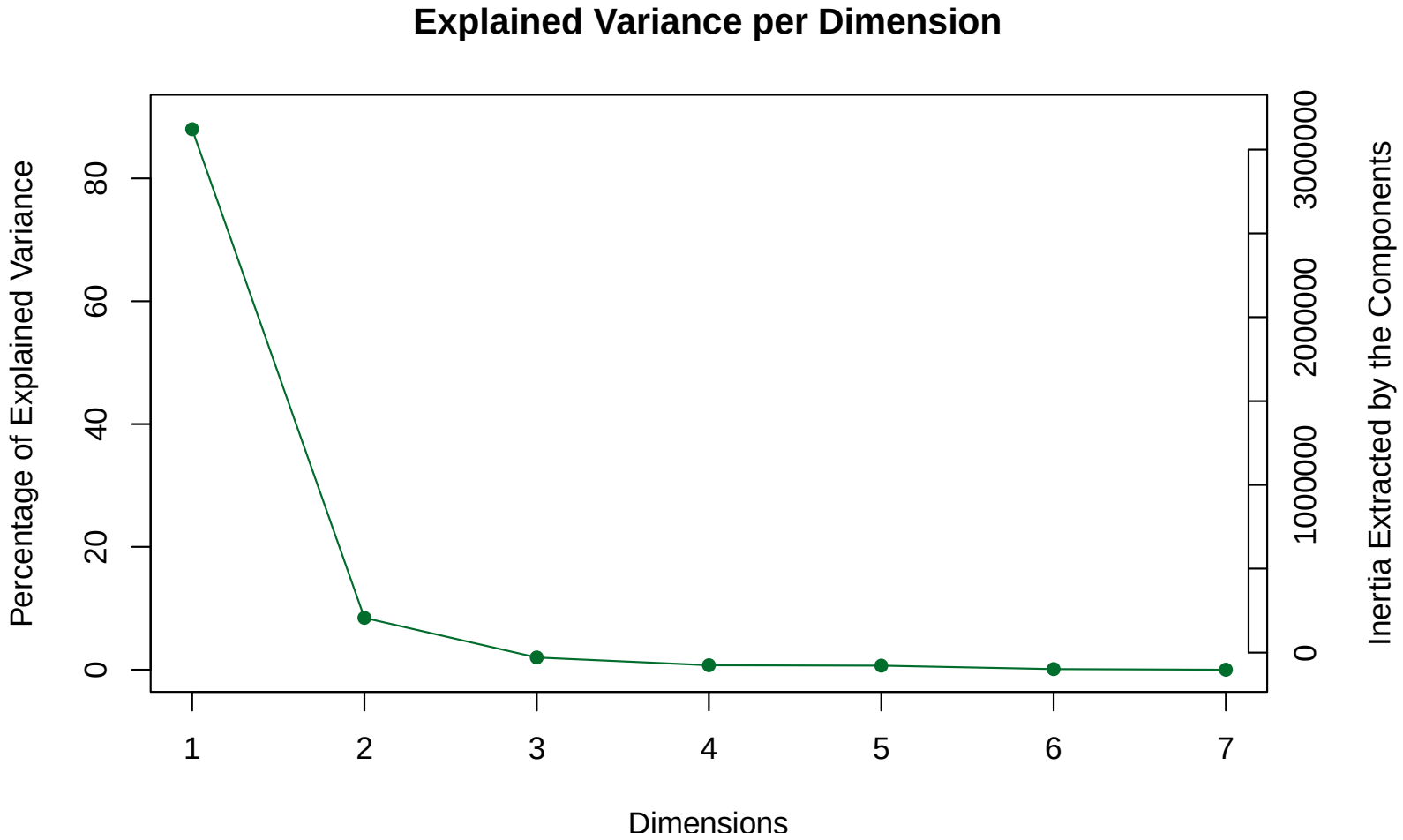
# Better: Covariance



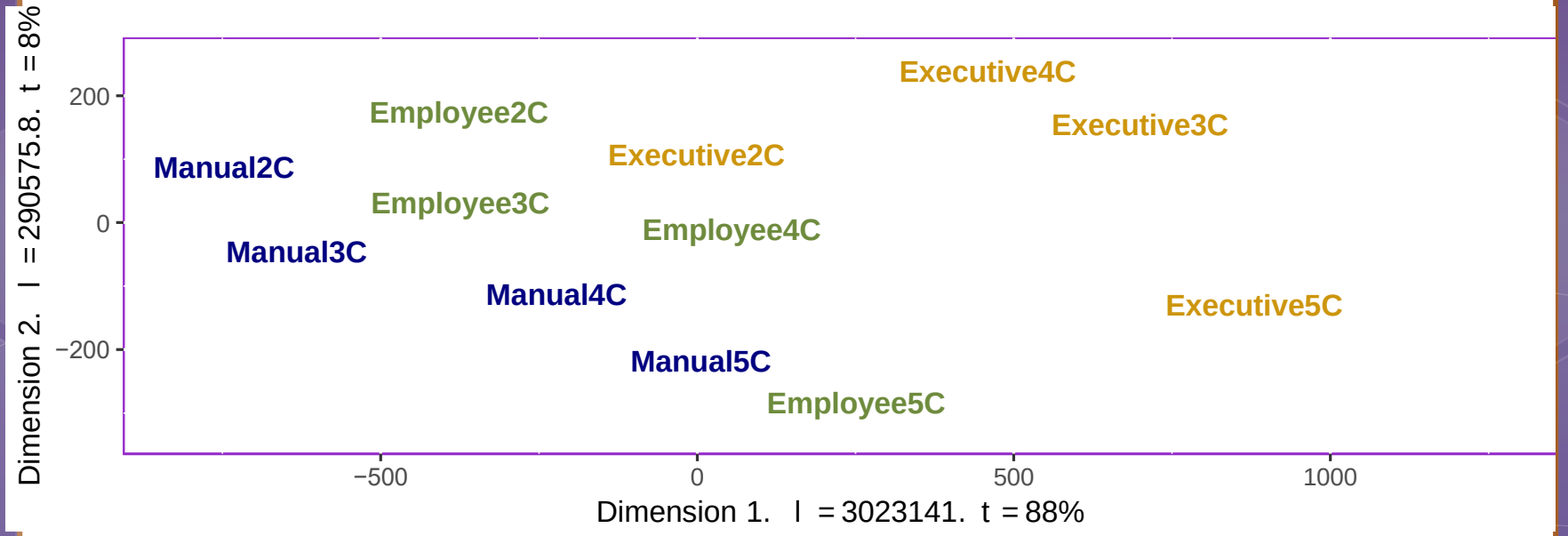
# Correlation

# PCA

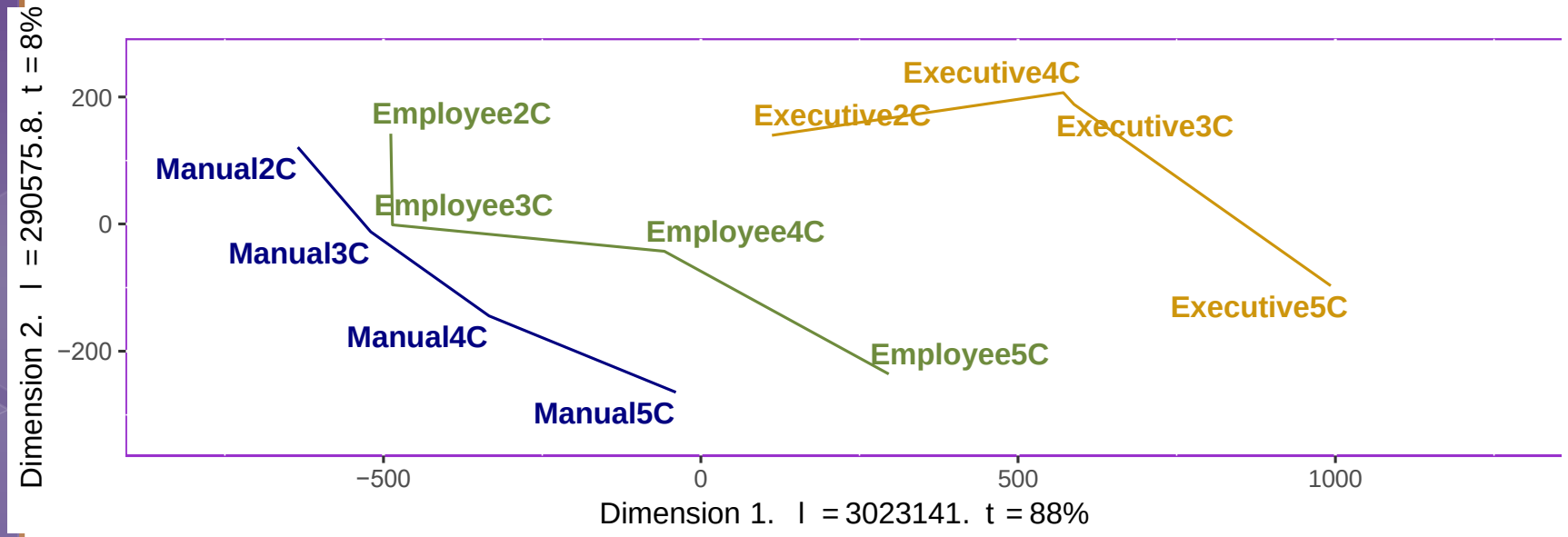
## RESPCA = EPPCA(X, SCALE = FALSE)



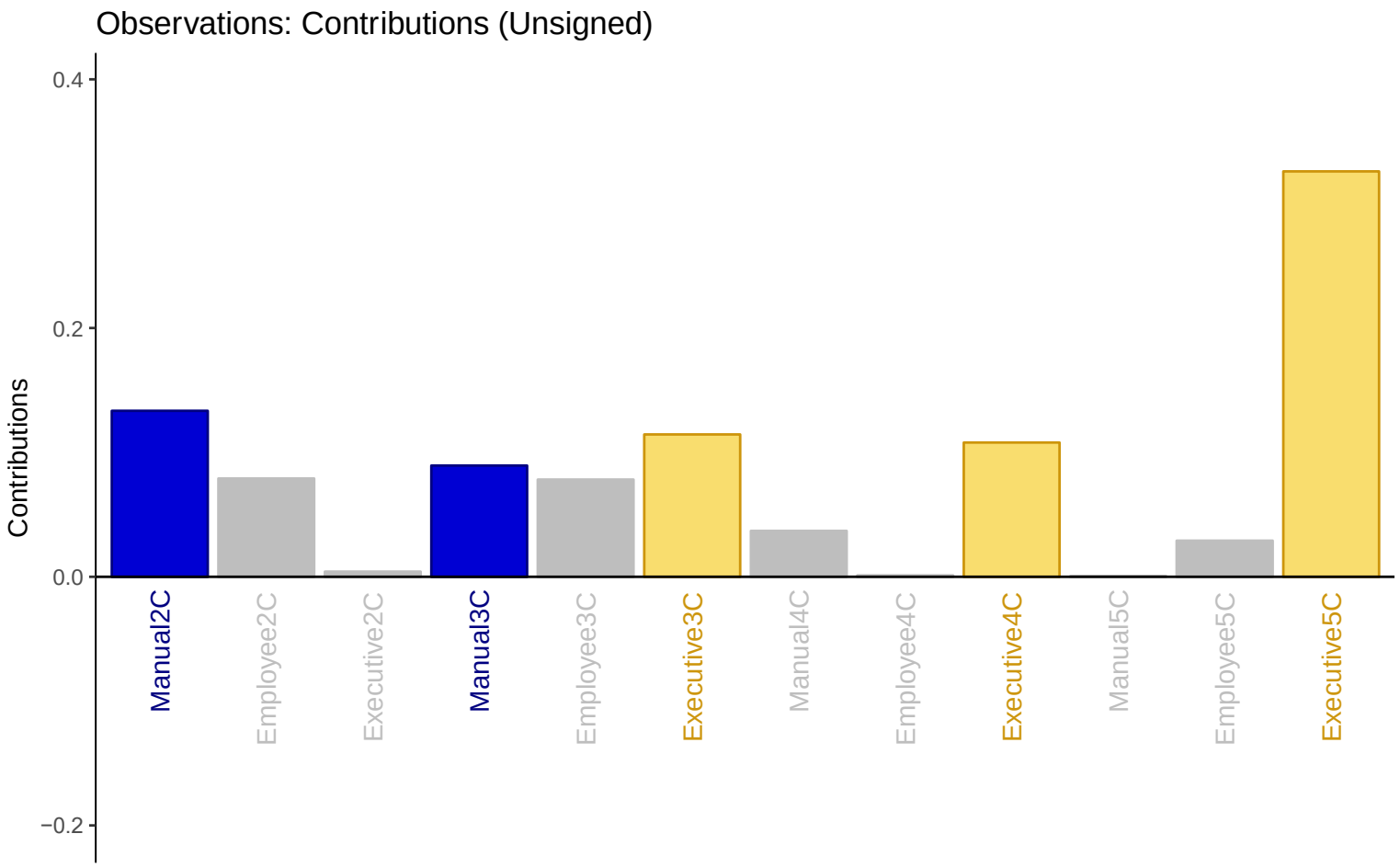
# Factor scores



# With lines

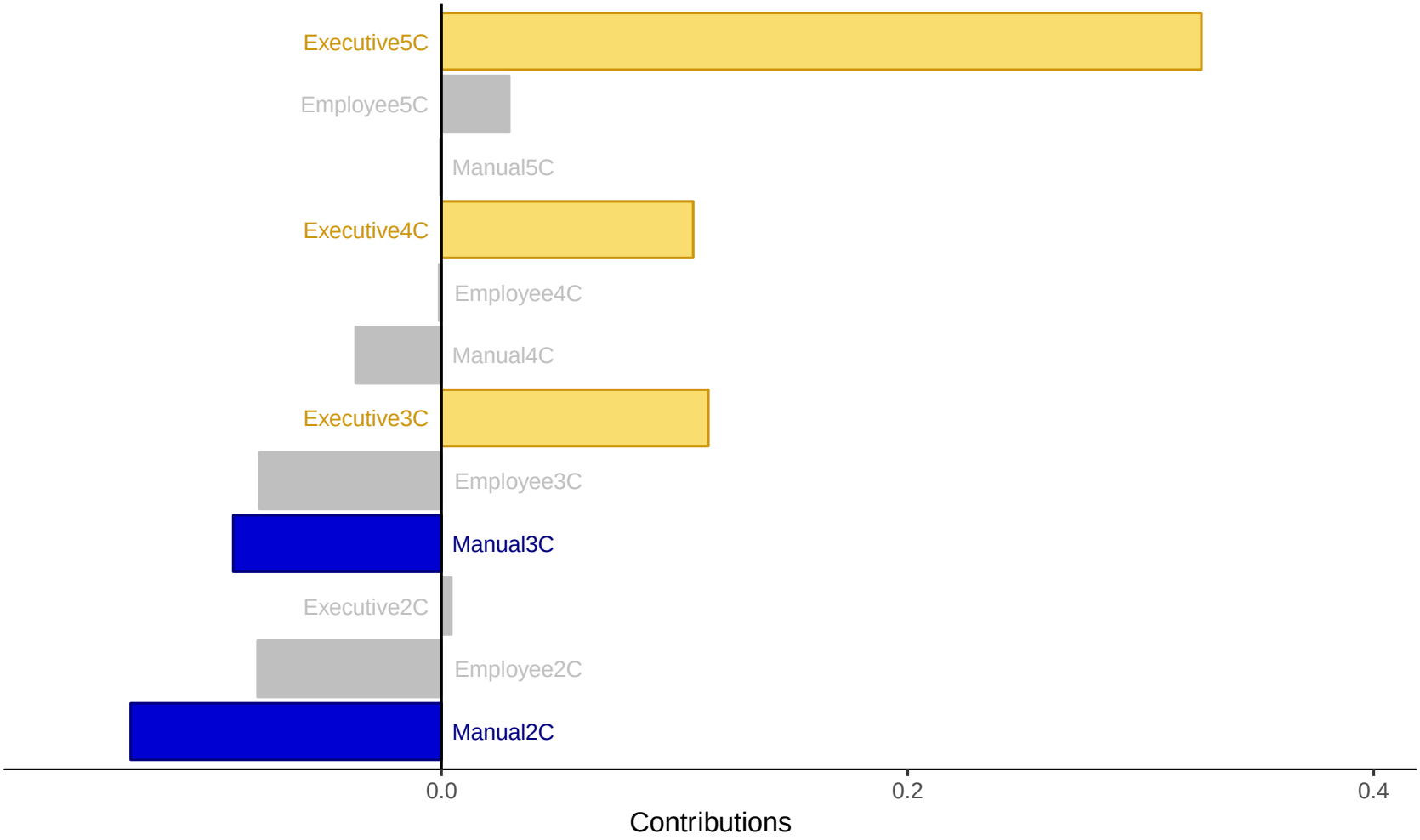

Executive4C
Executive2C
Executive3C
Employee2C
Manual2C
Employee3C
Employee4C
Manual3C
Executive5C
Manual4C
Employee5C
Manual5C
200
0
−200
−500
0
500
1000
Dimension 2. I = 290575.8. t = 8%
Dimension 1. I = 3023141. t = 88%

# Contributions

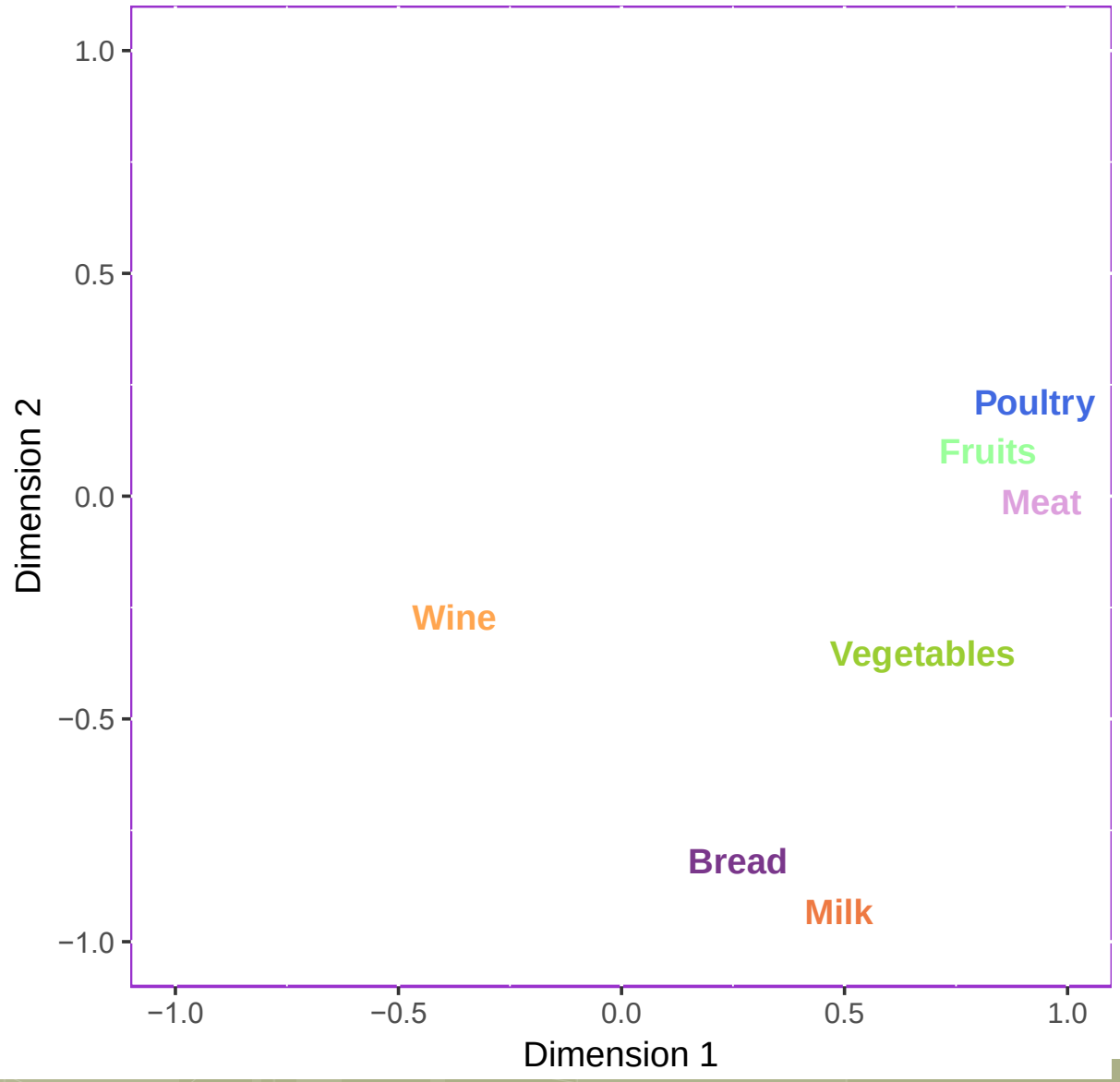


# Signed

Observations: Contributions (Signed)

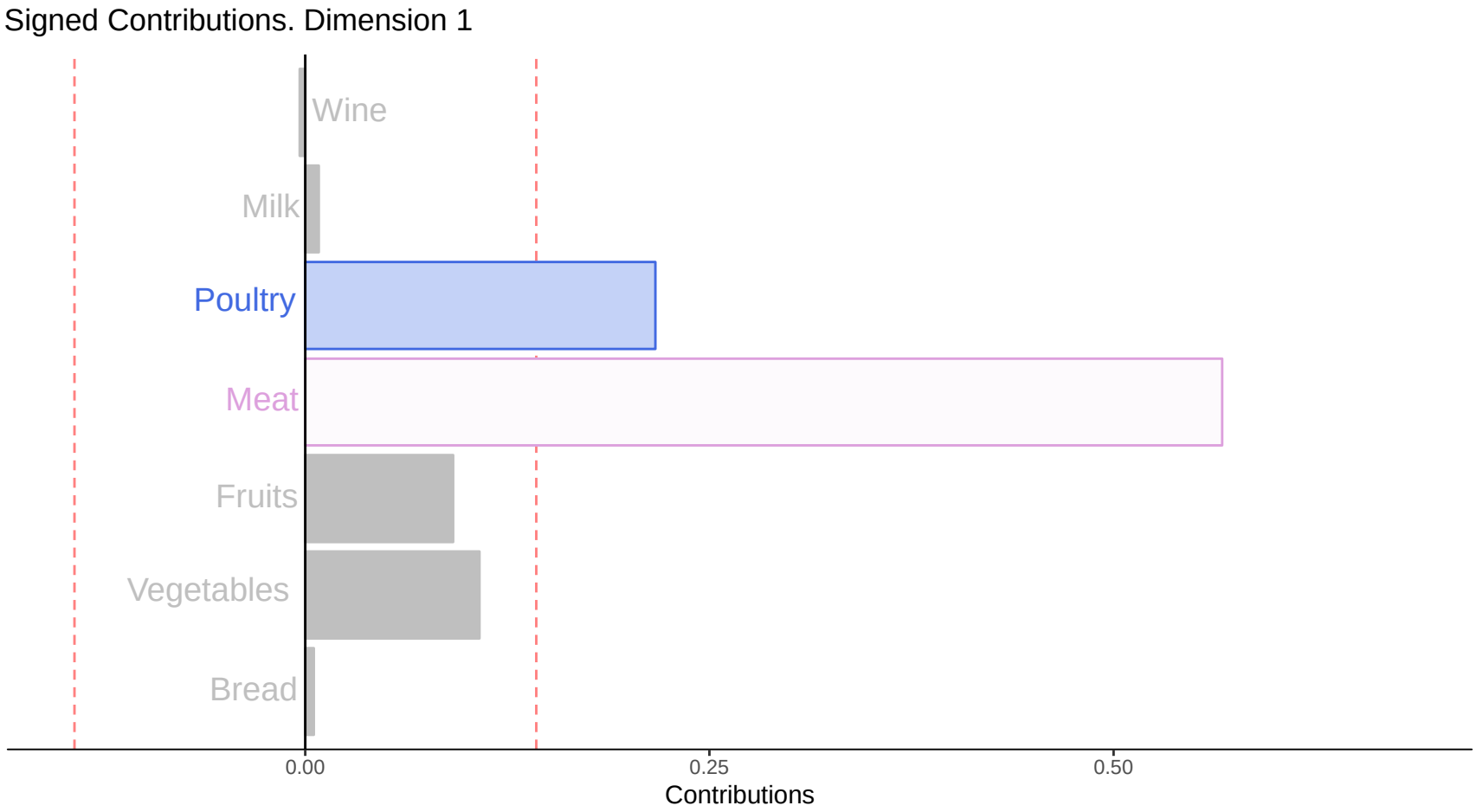


# The J-set


1.0
0.5
0.0
−0.5
−1.0
Dimension 2
−1.0
−0.5
0.0
0.5
1.0
Dimension 1
Poultry
Fruits
Meat
Wine
Vegetables
Bread
Milk

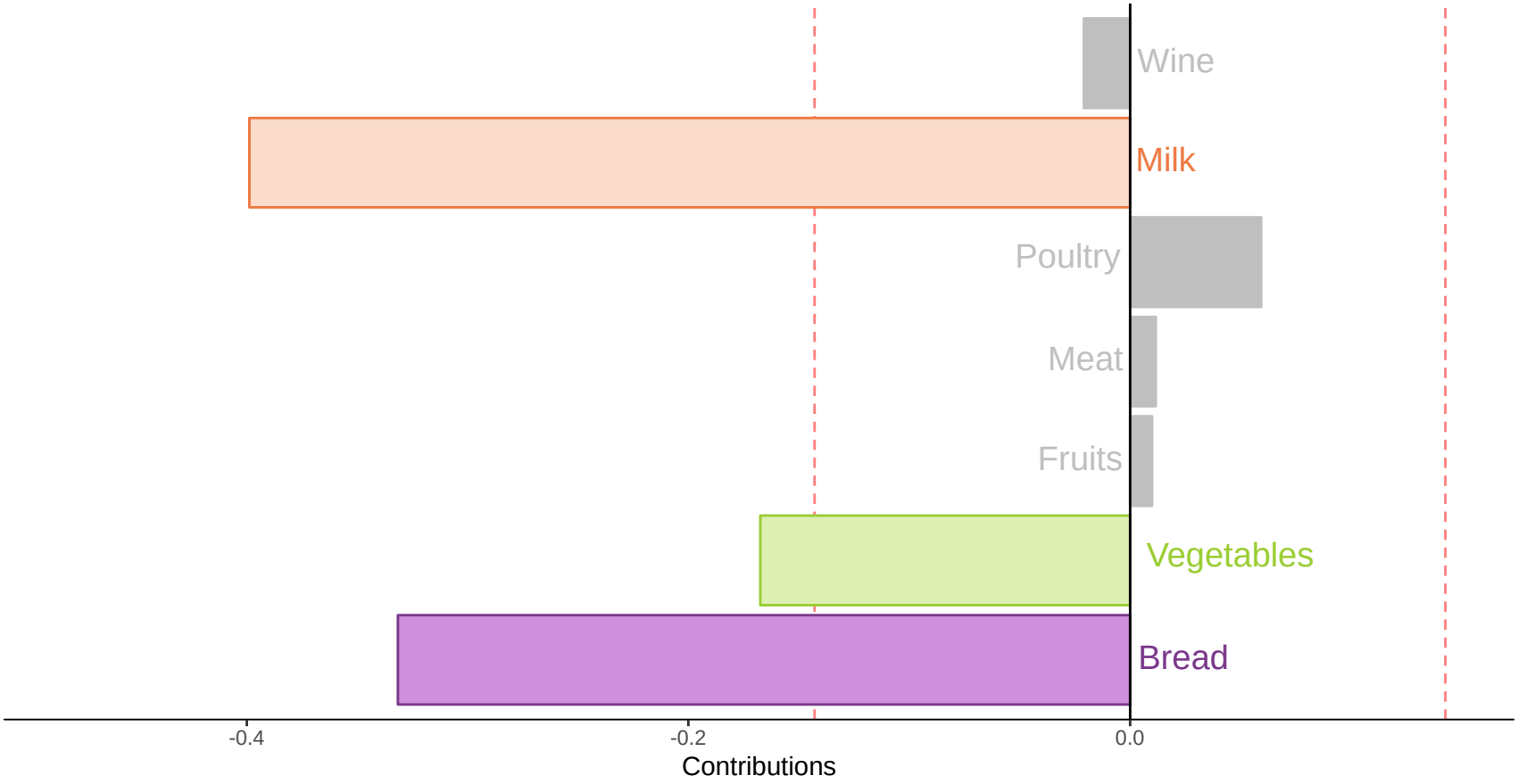
# What about the contributions

# DIMENSION 1.


Signed Contributions. Dimension 1
Wine
Milk
Poultry
Meat
Fruits
Vegetables
Bread
0.00
0.25
0.50
Contributions

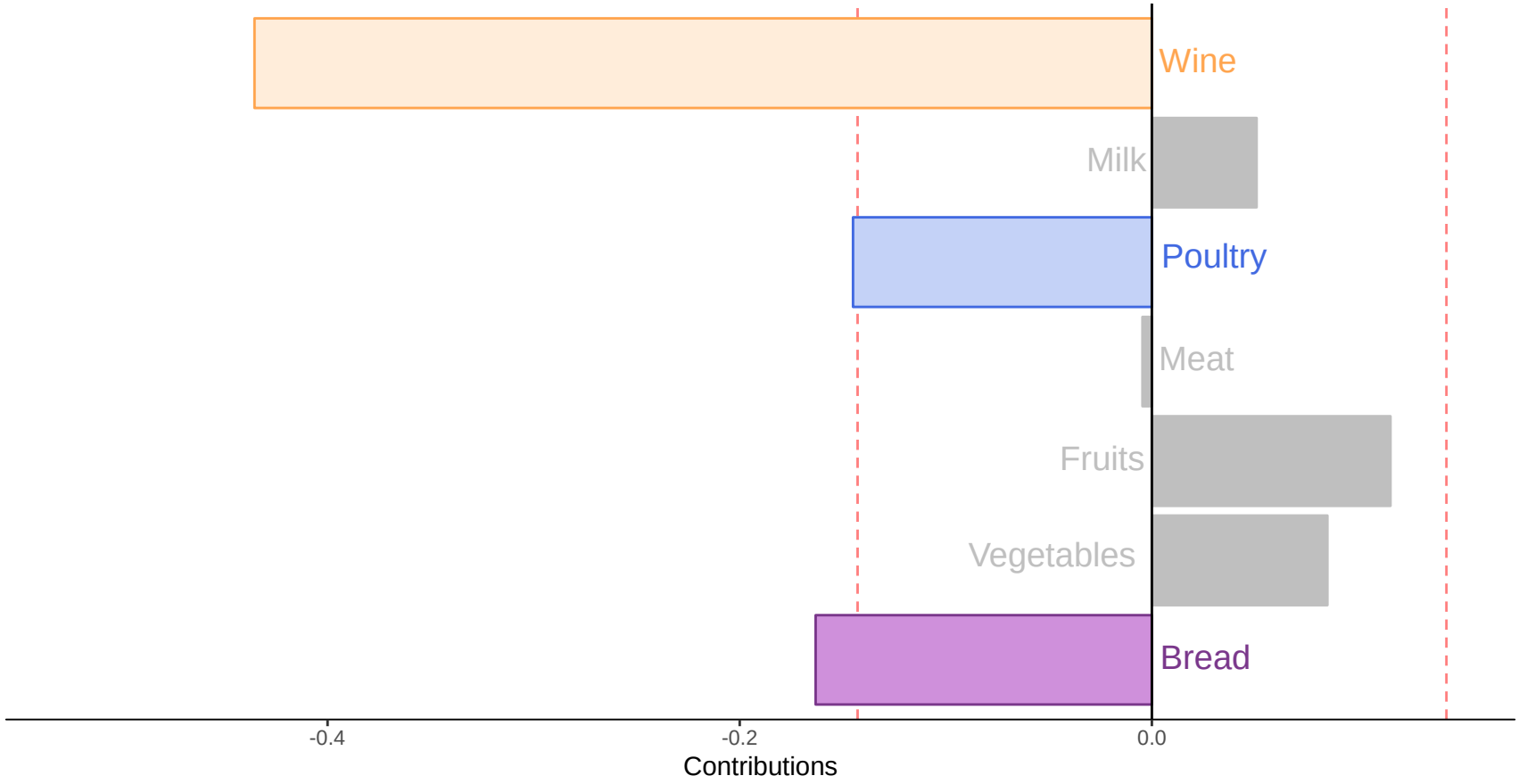
# DIMENSION 2.

Signed Contributions. Dimension 2



# Dimension 3

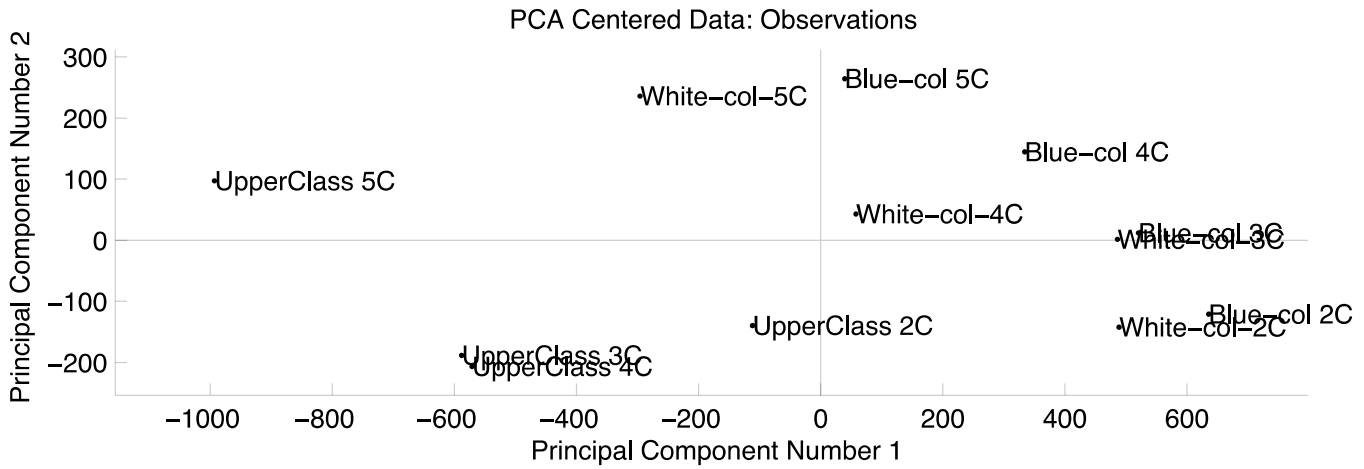
Signed Contributions. Dimension 3

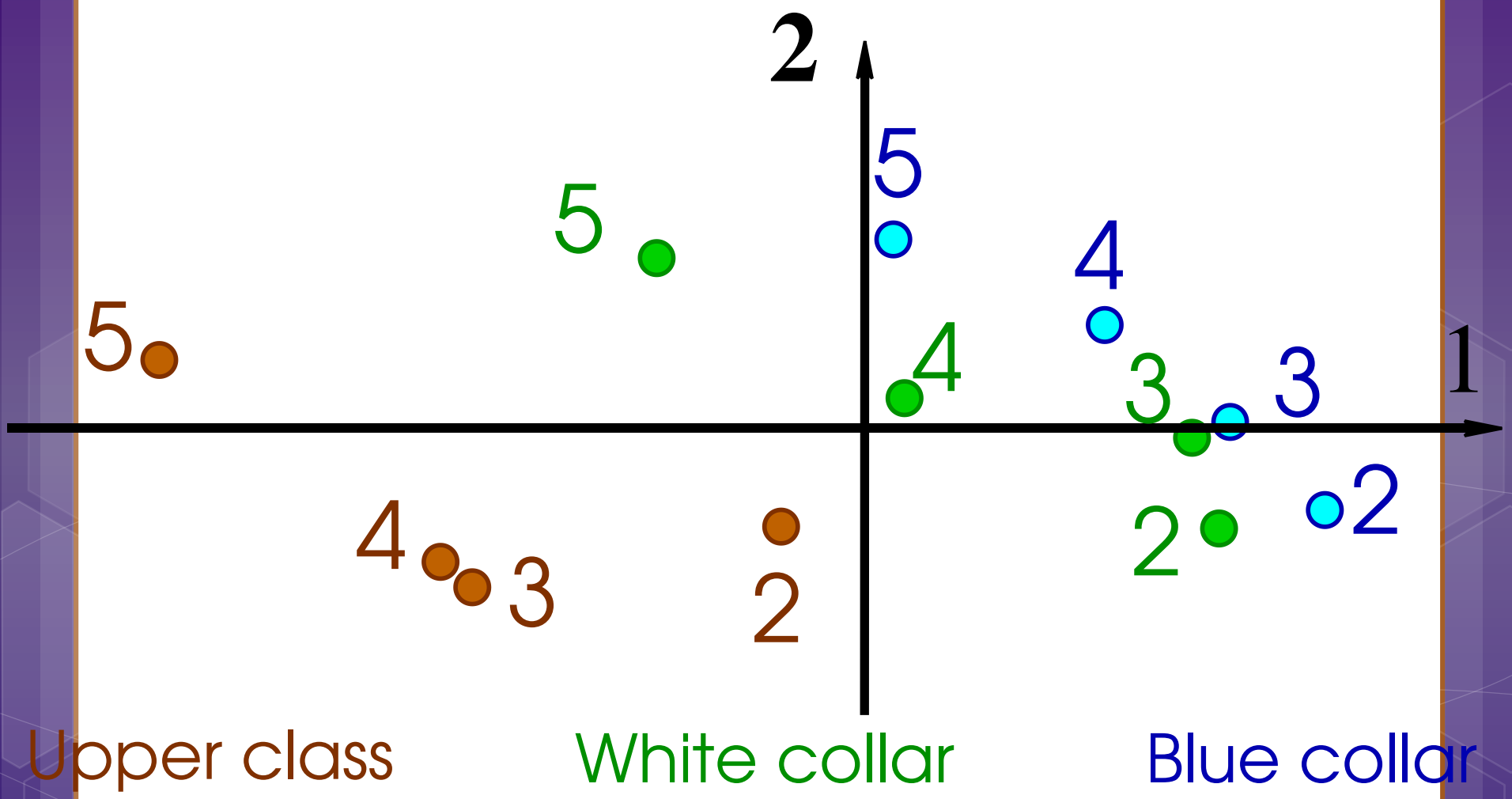


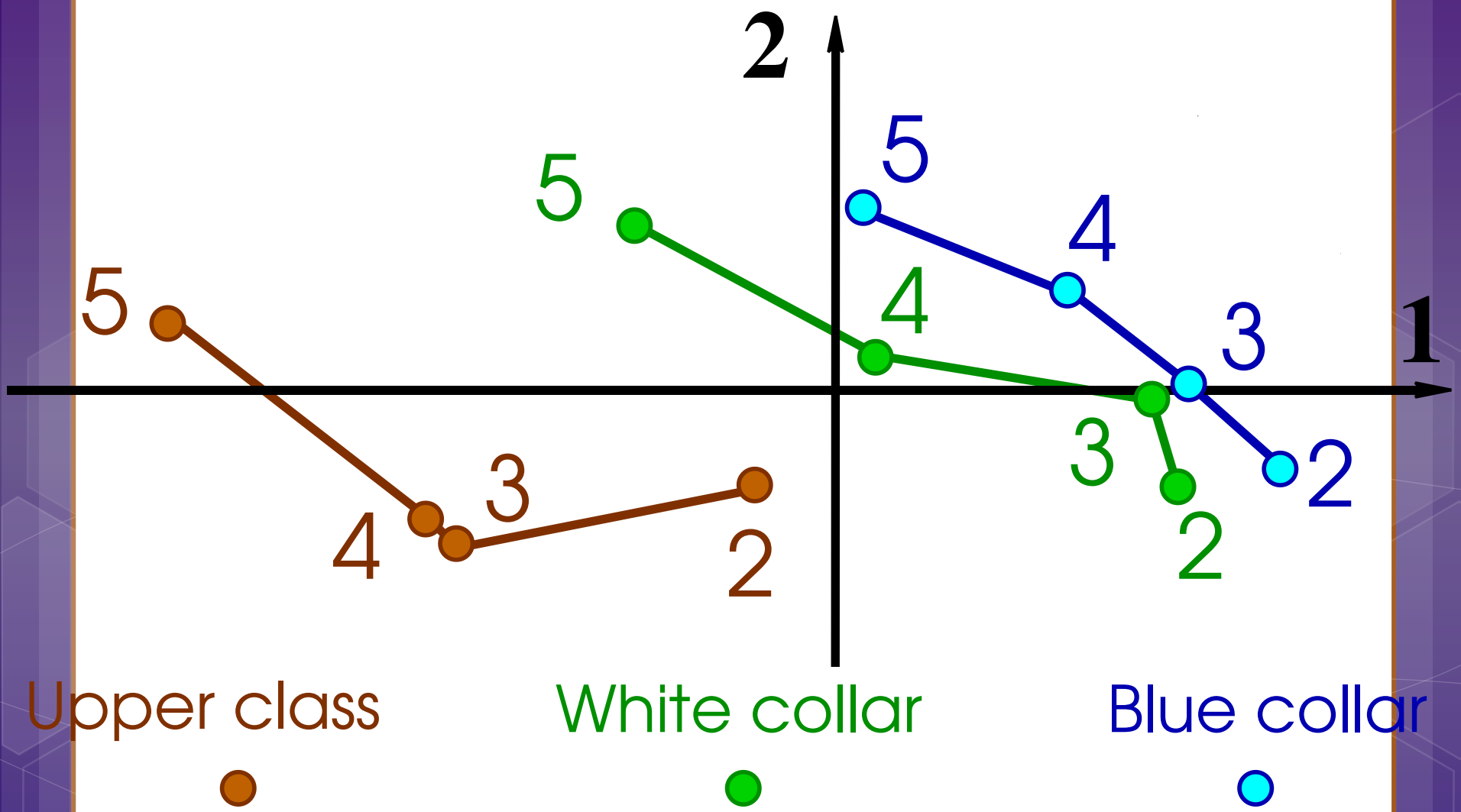
# Conclusions Food

# PCA. Analysis with MATLAB +

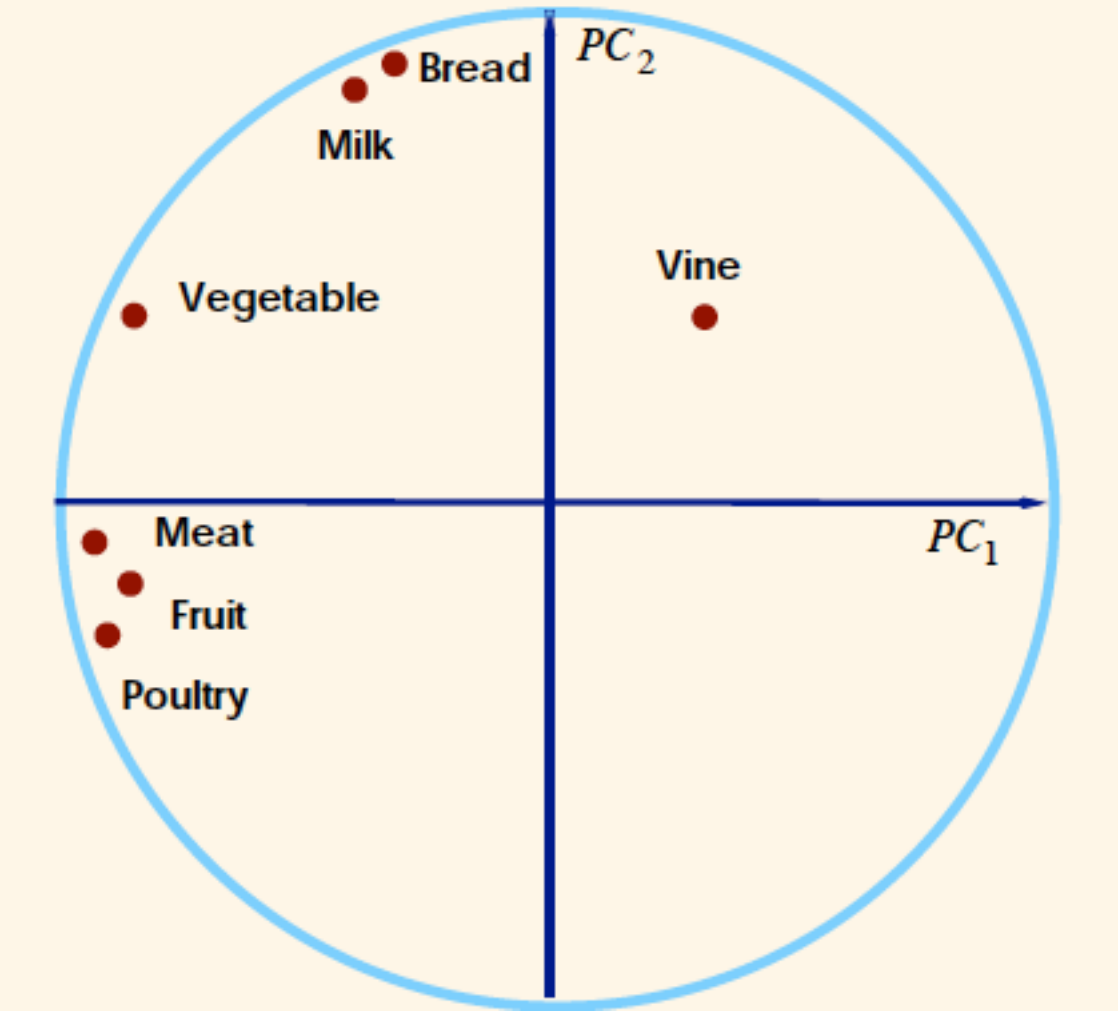
# NEVER DO THAT: THE HORRIBLE WAY




2
1
5
5
4
5
4
3
3
2
2
4
3
2
Upper class
White collar
Blue collar

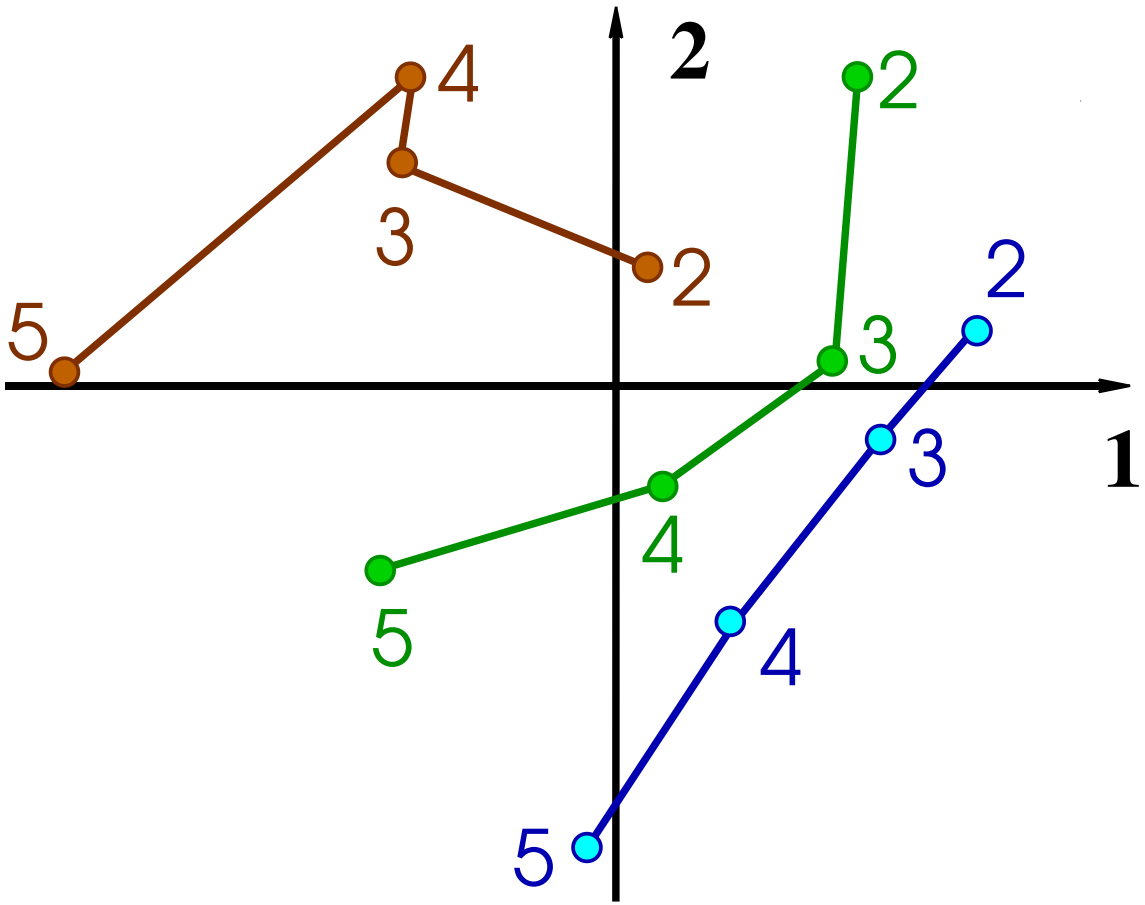

2
1
5
4
3
2
5
4
3
2
5
4
3
2
Upper class
White collar
Blue collar

# THE VARIABLES

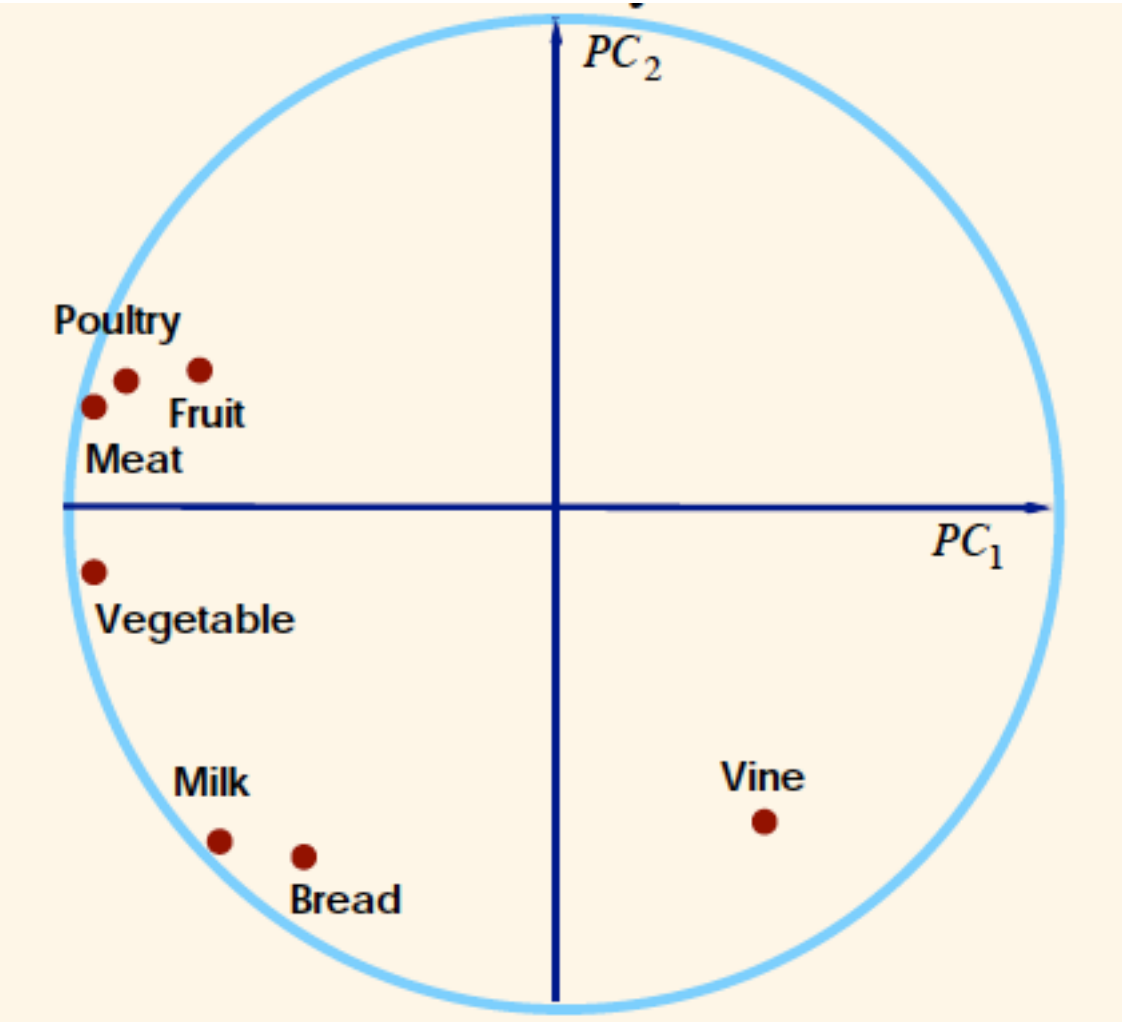


# The Z story

# The Z=story



# THE Z-STORY



# Important variables: Take II

# The great divide: Rotation

**Why? Psychometric Tradition**

- **Advocates: Thurstone (1947) and Cattell (1978)**
- **simplifies the factor structure**
- **facilitates interpretation**
- **increases reliability / replicability (with *dif- ferent* data samples).**

# Simple structure: Thurstone's 5 Criteria

1. Each row contains at least one zero

2. For each column, there are at least as many zeros as there are columns (i.e., number of factors kept)

3. For any pair of factors, there are some variables with zero loadings on one factor and large loadings on the other factor;

4. For any pair of factors, there is a sizable proportion of zero loadings

5. For any pair of factors, there is only a small number of large loadings

# How to do

**First: Choose the number of Factors to Keep**

**Second: Rotate to match a criterion**

# Orthogonal Rotations

- Most of Rotations is Varimax
  - maximizes the variance of the loadings
- Quartimax
  - minimizes number of factors per variable
- Equimax
  - combine varimax and quartimax

# Oblique Rotations

- **Promax: a 2 step procedure.**

  **Goal: create binary factors**

  **1. Create ?target matrix:? Start with Varimax and raise to power 3 or 4 (or more). This creates binary factors.**

  **2. Fit the original factors to the target with procruste analysis (What is that?).**

- **Newcomer:**
  - **Independent Component Analysis (ICA)**

# New Ideas: Sparsification and Lasso

# When to rotate

- French School: Never!
- American school: Always!
- So, the question is:
  - When does rotating make sense?
  - When the Psychometric model is valid!

# SILLY EXAMPLE FOR ROTATION

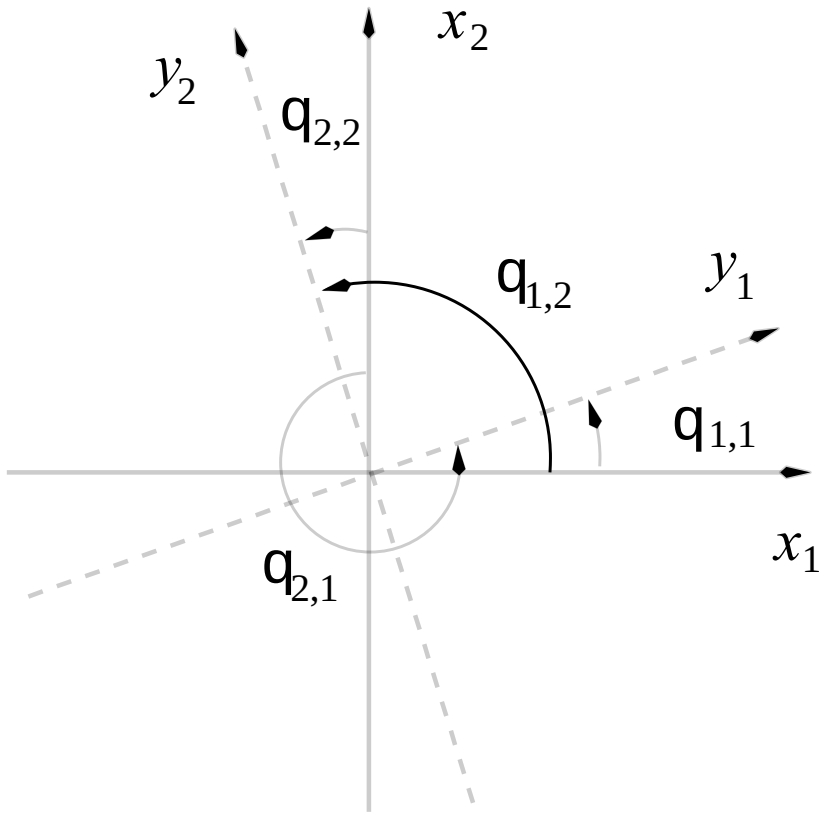


Table 10: An (artificial) example for PCA and rotation. Five wines are described by seven variables.

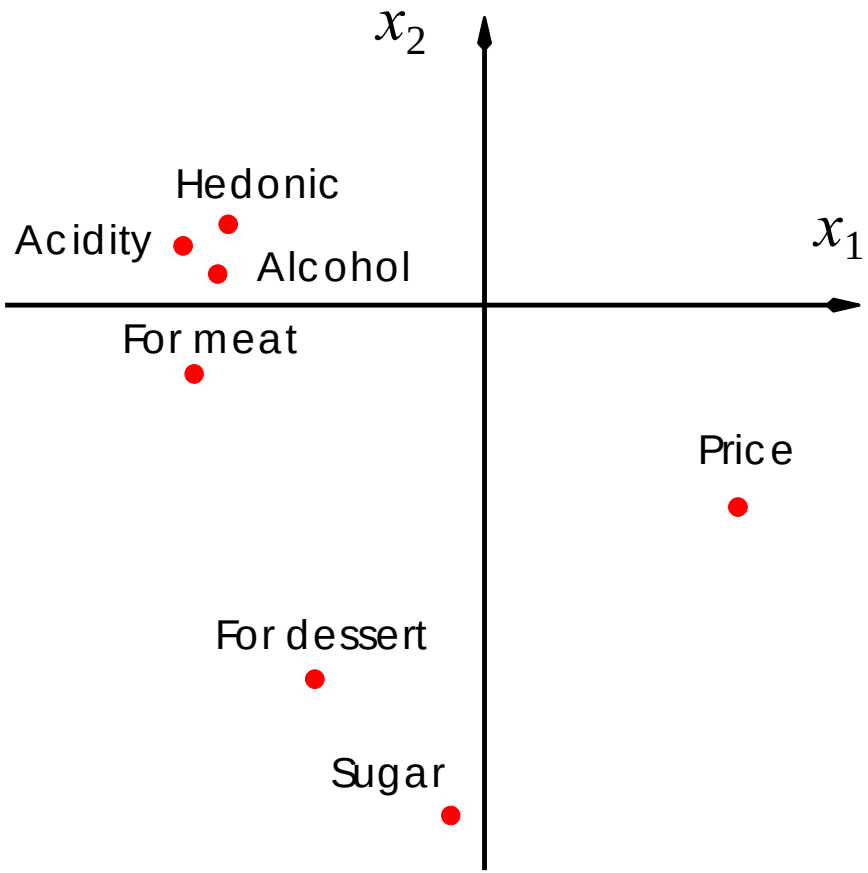
| | Hedonic | For meat | For dessert | Price | Sugar | Alcohol | Acidity |
|---|---|---|---|---|---|---|---|
| Wine 1 | 14 | 7 | 8 | 7 | 7 | 13 | 7 |
| Wine 2 | 10 | 7 | 6 | 4 | 3 | 14 | 7 |
| Wine 3 | 8 | 5 | 5 | 10 | 5 | 12 | 5 |
| Wine 4 | 2 | 4 | 7 | 16 | 7 | 11 | 3 |
| Wine 5 | 6 | 2 | 4 | 13 | 3 | 10 | 3 |

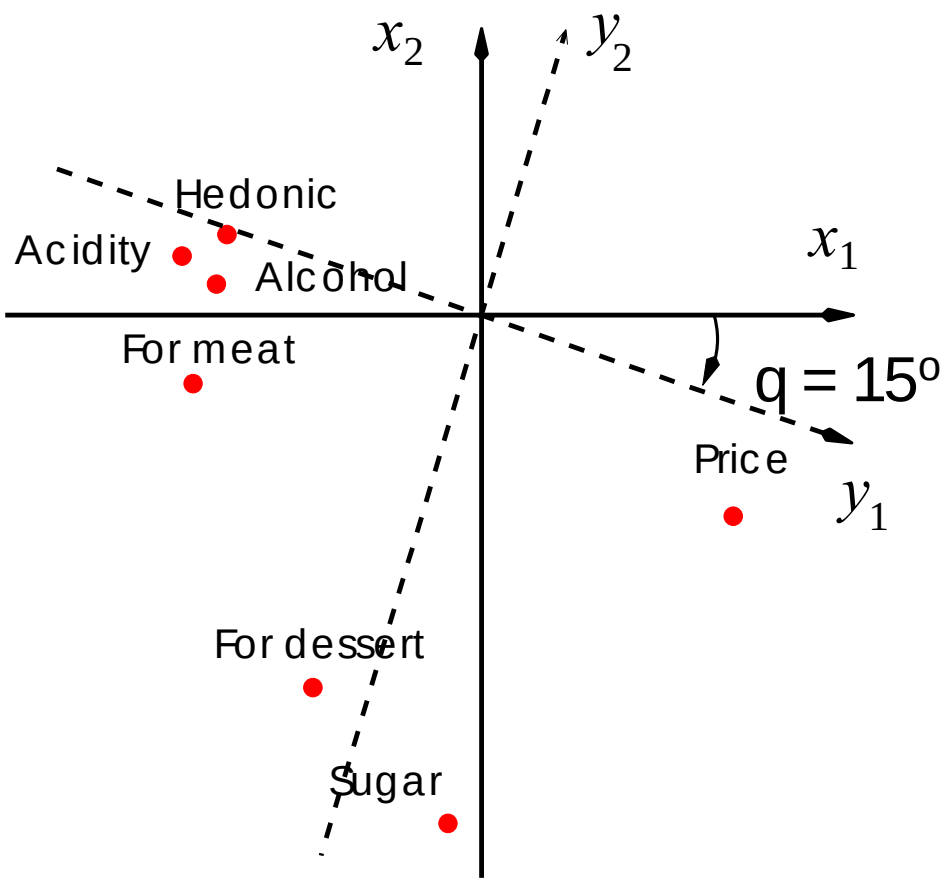
4 Factors. (Eigenvalues of 4.7627, 1.8101, 0.3527, 0.0744, respectively), 2 factors: 94% of the variance.

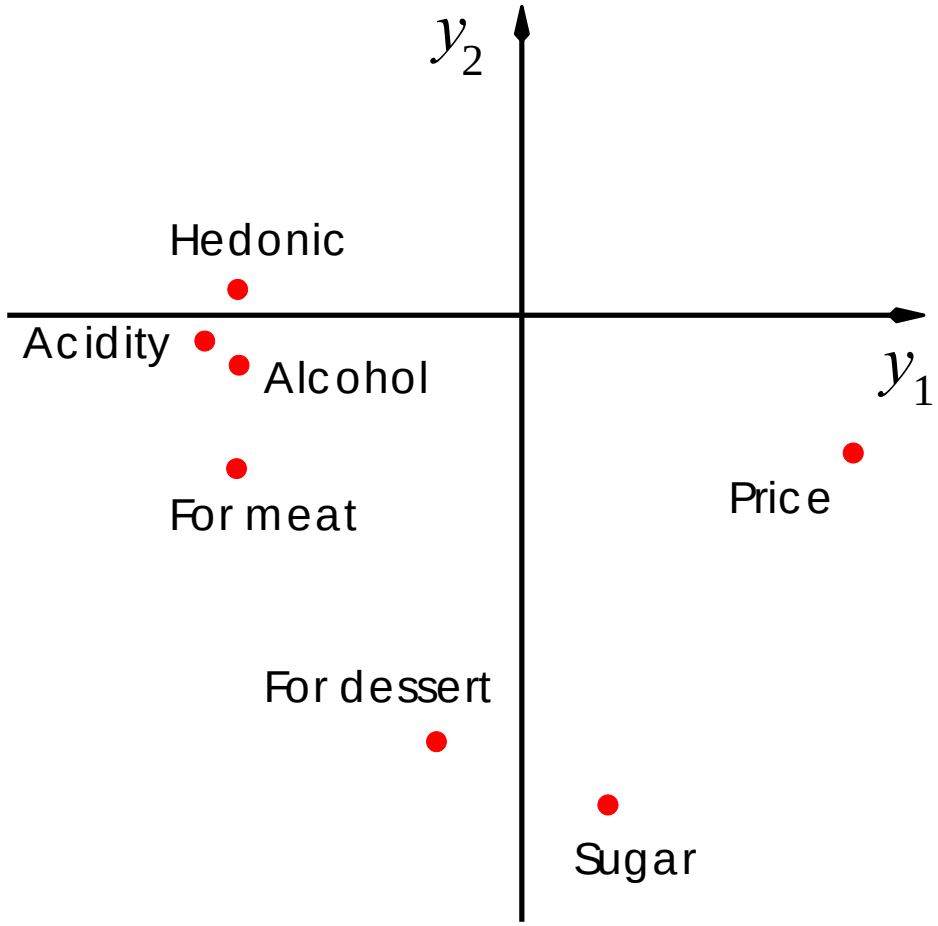
Table 11: Wine example: Original loadings of the seven variables on the first two components.

| | Hedonic | For meat | For dessert | Price | Sugar | Alcohol | Acidity |
|---|---|---|---|---|---|---|---|
| Factor 1 | −0.3965 | −0.4454 | −0.2646 | 0.4160 | −0.0485 | −0.4385 | −0.4547 |
| Factor 2 | 0.1149 | −0.1090 | −0.5854 | −0.3111 | −0.7245 | 0.0555 | 0.0865 |

# ORIGINAL LOADINGS




$x_2$
$y_2$
Hedonic
Acidity
Alcohol
$x_1$
For meat
q = 15º
Price
$y_1$
For dessert
Sugar


$y_2$
Hedonic
Acidity
Alcohol
For meat
$y_1$
Price
For dessert
Sugar

# Rotation: A better example

## Mini-Osiq: 14 out of 30

# ROTATIONS?

### Spatial 1, 9, 13, 14, 17, 22, 25, 28, 30

I was very good in 3-D geometry as a student.

I can easily imagine and mentally rotate 3-dimensional geometric figures.

I can easily sketch a blueprint for a building that I am familiar with.

I am a good Tetris player.

I have excellent abilities in technical graphics.

I find it difficult to imagine how a 3-dimensional geometric figure

My graphic abilities would make a career in architecture relatively easy for me.

### Object 4, 12, 17, 22, 25, 28, 30,

My images are very colourful and bright.

My images are very vivid and photographic.

When I imagine the face of a friend, I have a perfectly clear and bright image.

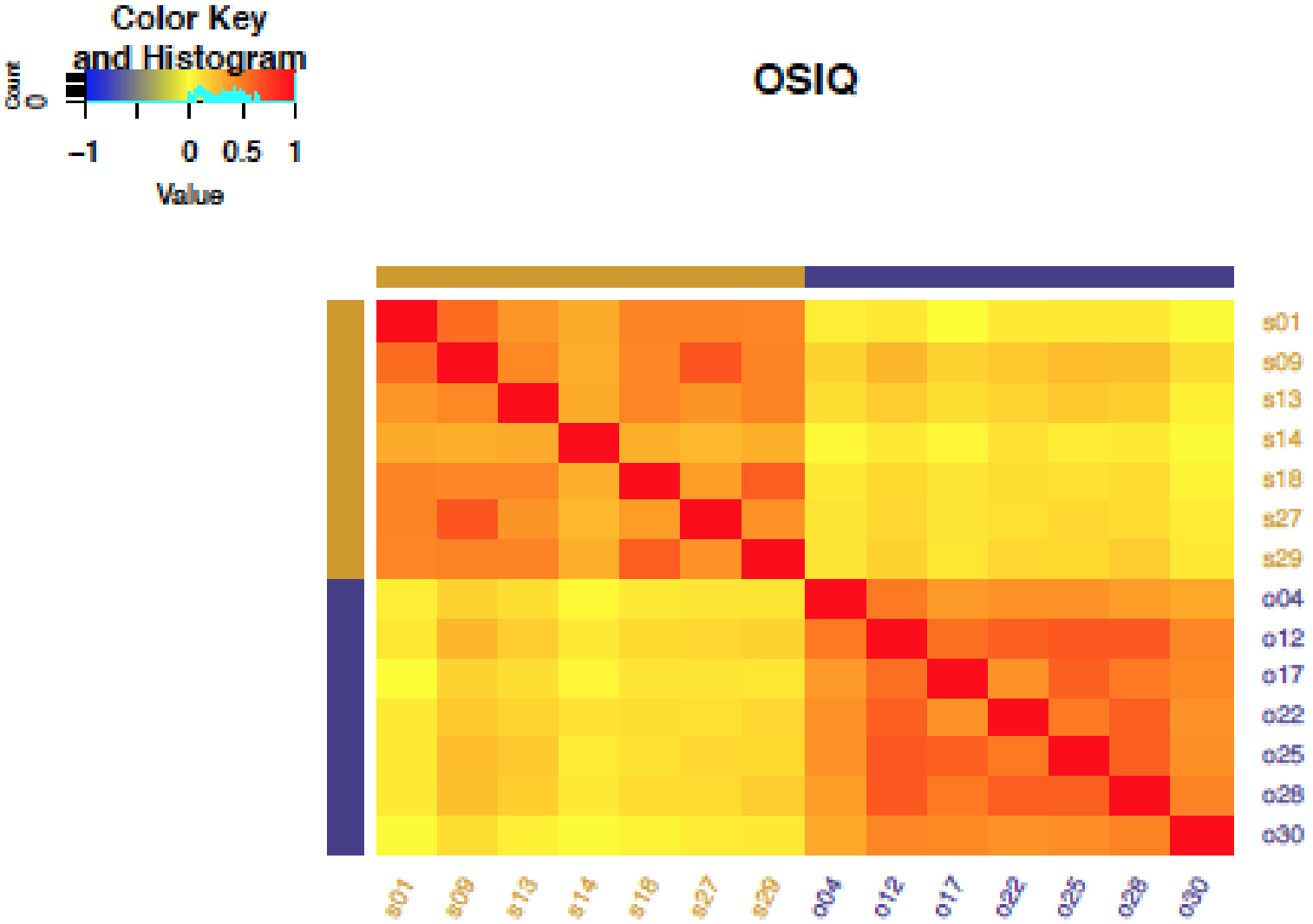
Sometimes my images are so vivid and persistent that it is difficult to ignore them.

I can close my eyes and easily picture a scene that I have experienced.

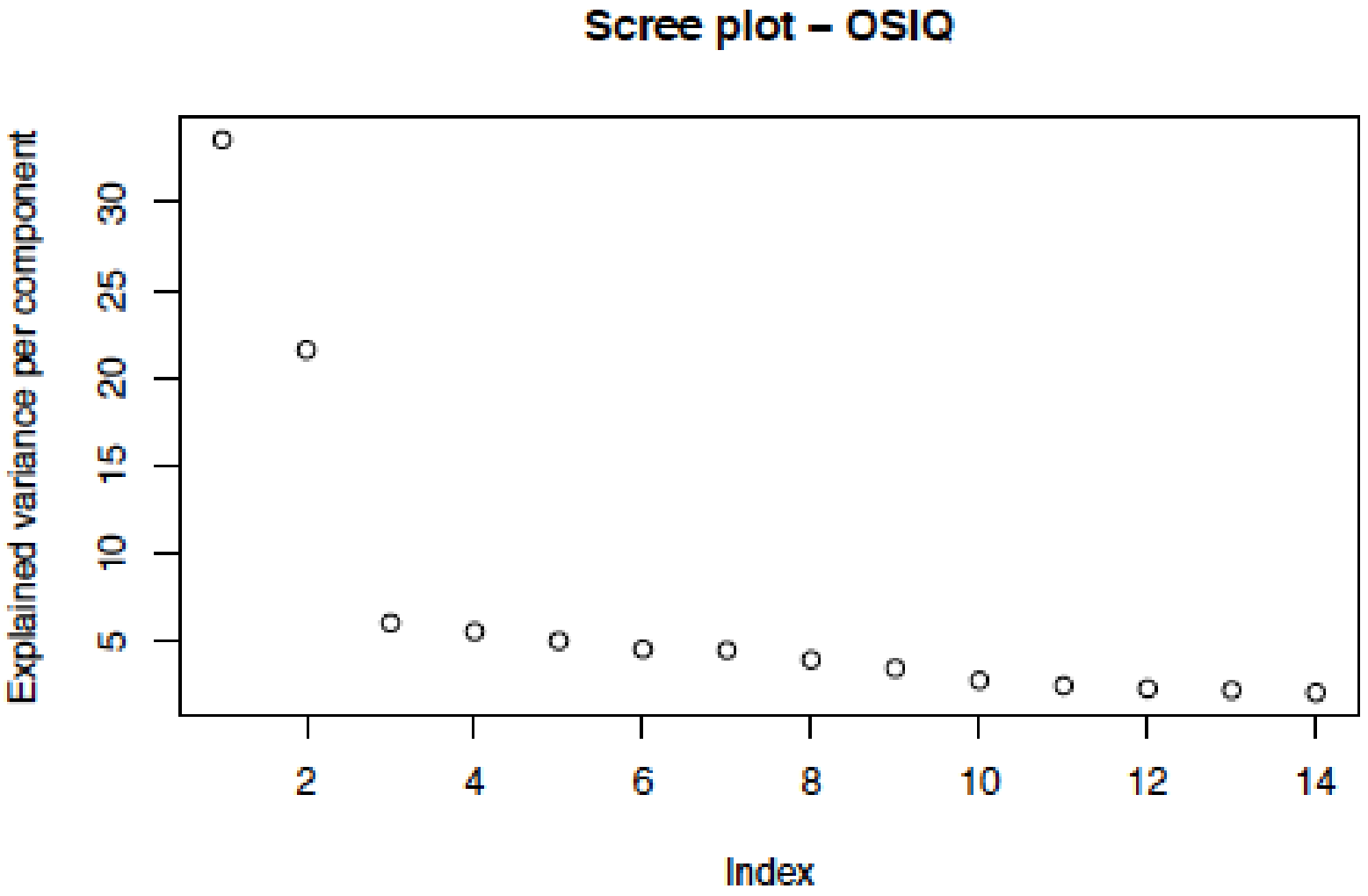
My visual images are in my head all the time. They are just right there.

When I hear a radio announcer or a DJ I've never actually seen, I usually find myself picturing what he or she might look like.

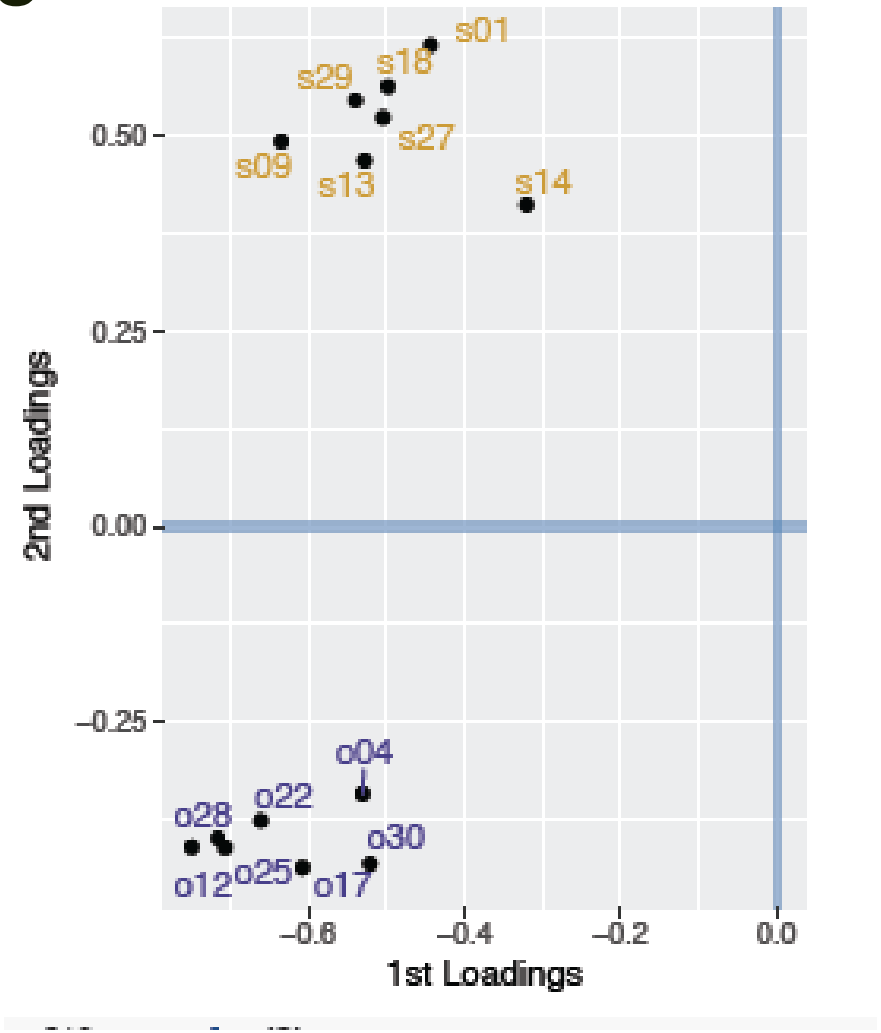
# A Heat Map


Color Key
and Histogram
Count
0
−1
0
0.5
1
Value
OSIQ
s01
s09
s13
s14
s18
s27
s29
o04
o12
o17
o22
o25
o28
o30

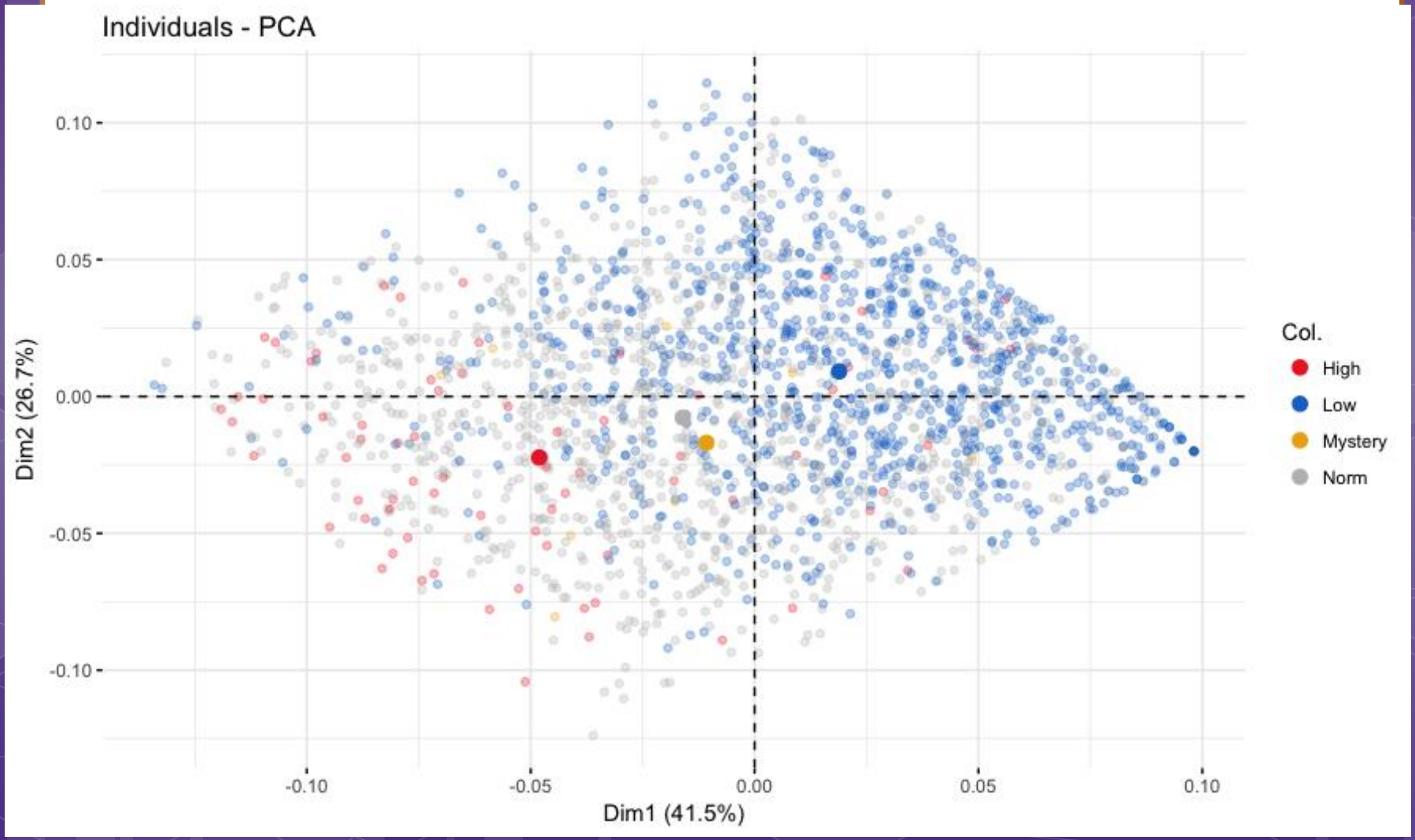
# SCREE

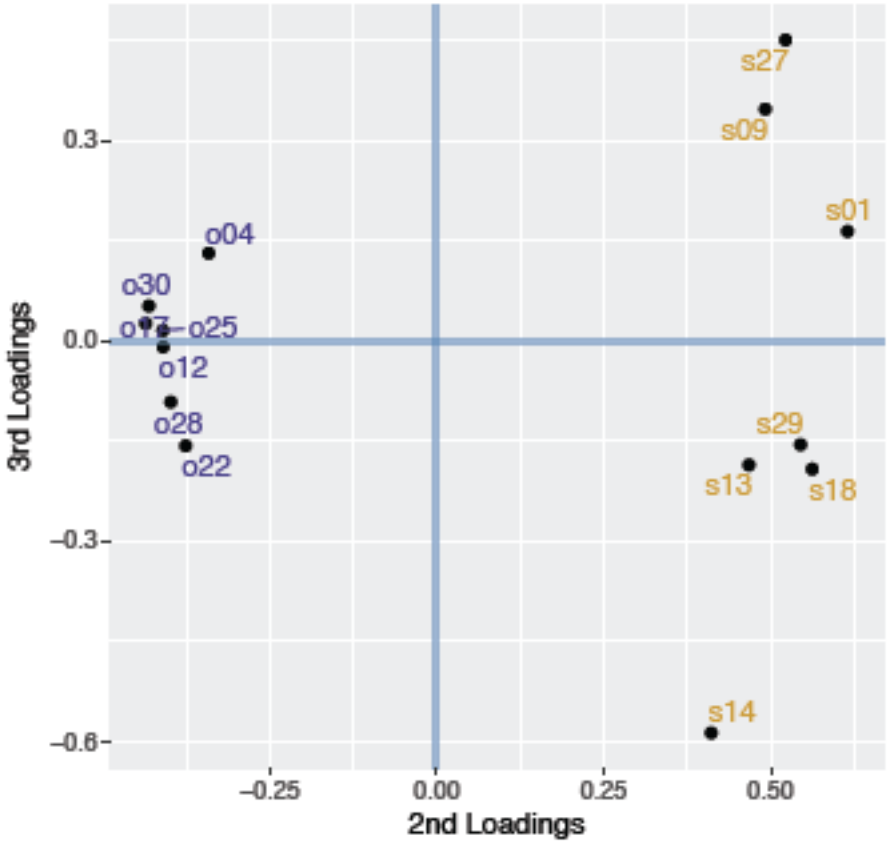

Scree plot – OSIQ
Explained variance per component
5
10
15
20
25
30
2
4
6
8
10
12
14
Index

# Loadings


s01
s18
s29
s27
s09
s13
s14
o04
o22
o28
o30
o12
o25
o17
0.50
0.25
0.00
−0.25
2nd Loadings
−0.6
−0.4
−0.2
0.0
1st Loadings

# UNROTATED I-SET




s27
s09
s01
o04
o30
o25
o12
o28
o22
s29
s13
s18
s14
0.3
0.0
−0.3
−0.6
−0.25
0.00
0.25
0.50
3rd Loadings
2nd Loadings

# Varimax here

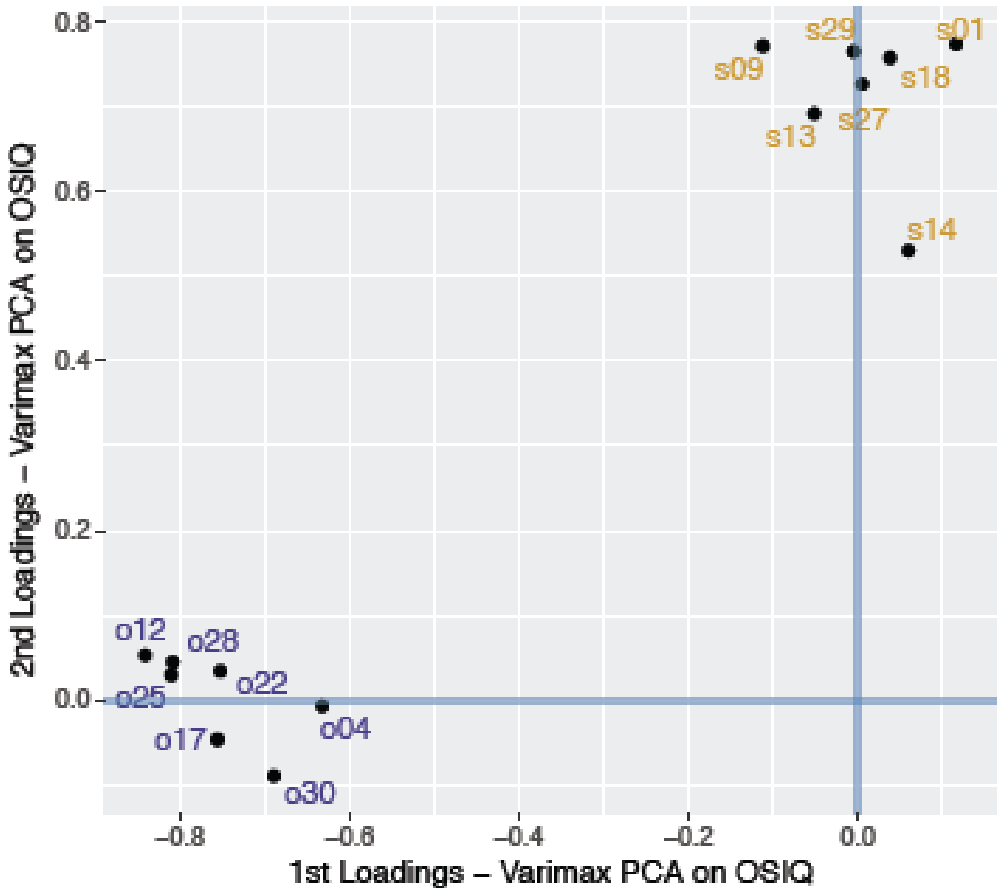
```
din.rot <- 2 # how many factors do we want to keep for rotation
Y_varimax = varimax(ResPCA_Y$ExPosition.Data$fj[,1:din.rot],
                                    normalize = TRUE, eps = 1e-5)
```

Rotation matrix from varimax:

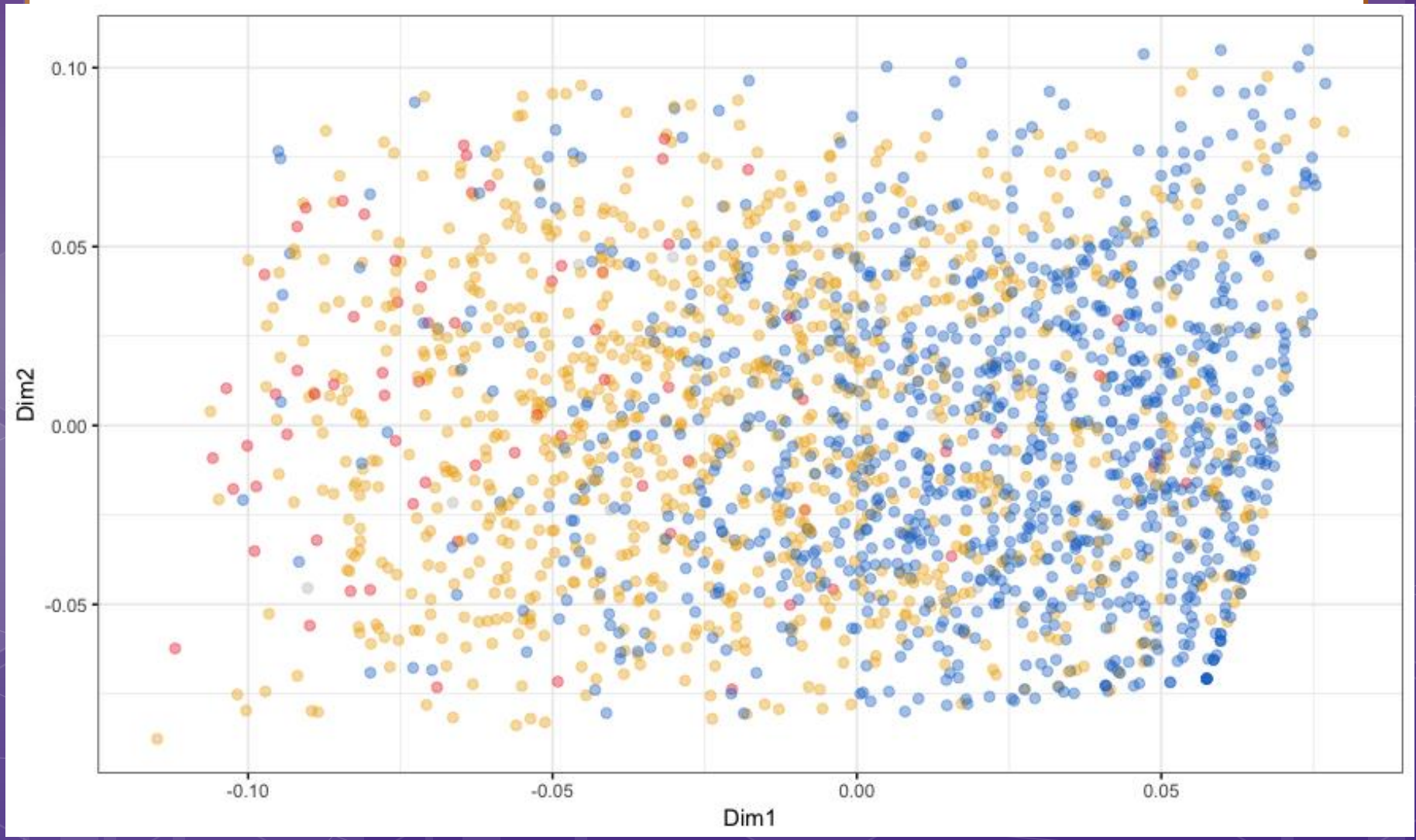
```
          [,1]       [,2]
[1,] 0.7868091 -0.6171964
[2,] 0.6171964  0.7868091
```

Rotation: arcos(.617) = 52 degrees

# VARIMAX ON LOADINGS



# ROTATED I-SET



# Important variables: Take III

# Inferences?

All that is descriptive.

# The earth is round

## The earth is round:

$p < .05$ (cf. Jacob Cohen, 1994)

## Are the results to be trusted?

## Question 1. Is there anything in the data?

## Question 2. How Robust are the results?

## Question 1.

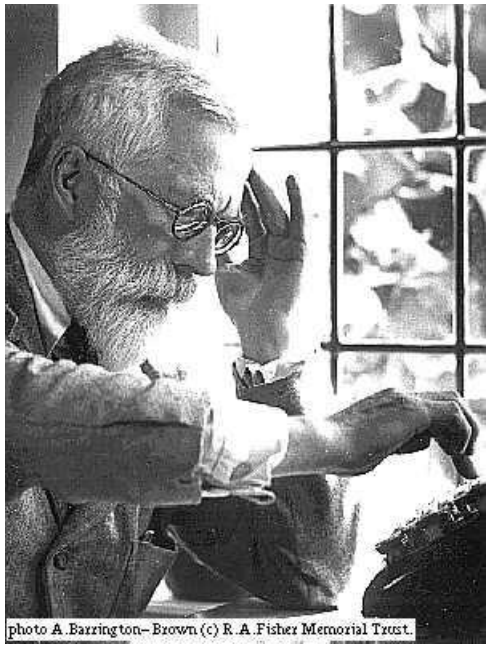
## Is there anything in the data

## Compare results to chance (nothing)

# Permutation Test

William S. Gosset *aka* Student
(1876-1937)

Ronald A. Fisher
(1890-1962)



# How to get chance?

## Put the data for Variable 1 in a Hat

# How to get chance?

## Put the data for Variable 1 in a Hat

# How to get chance?

## Put the data for Variable 1 in a Hat

# A Shake!

# How to get chance?

Get them out!

# How to get chance?

## Put the data for Variable 2 in a Hat

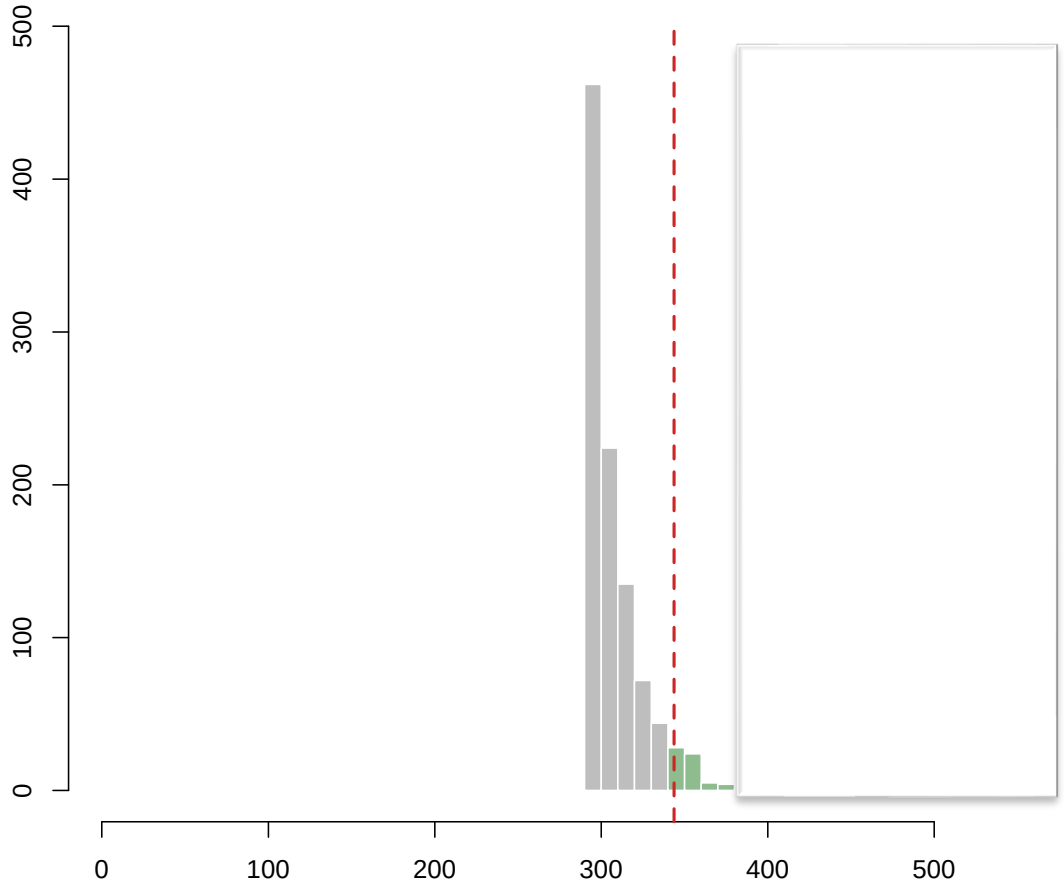
## Get them out!

# So the link between variables is broken!

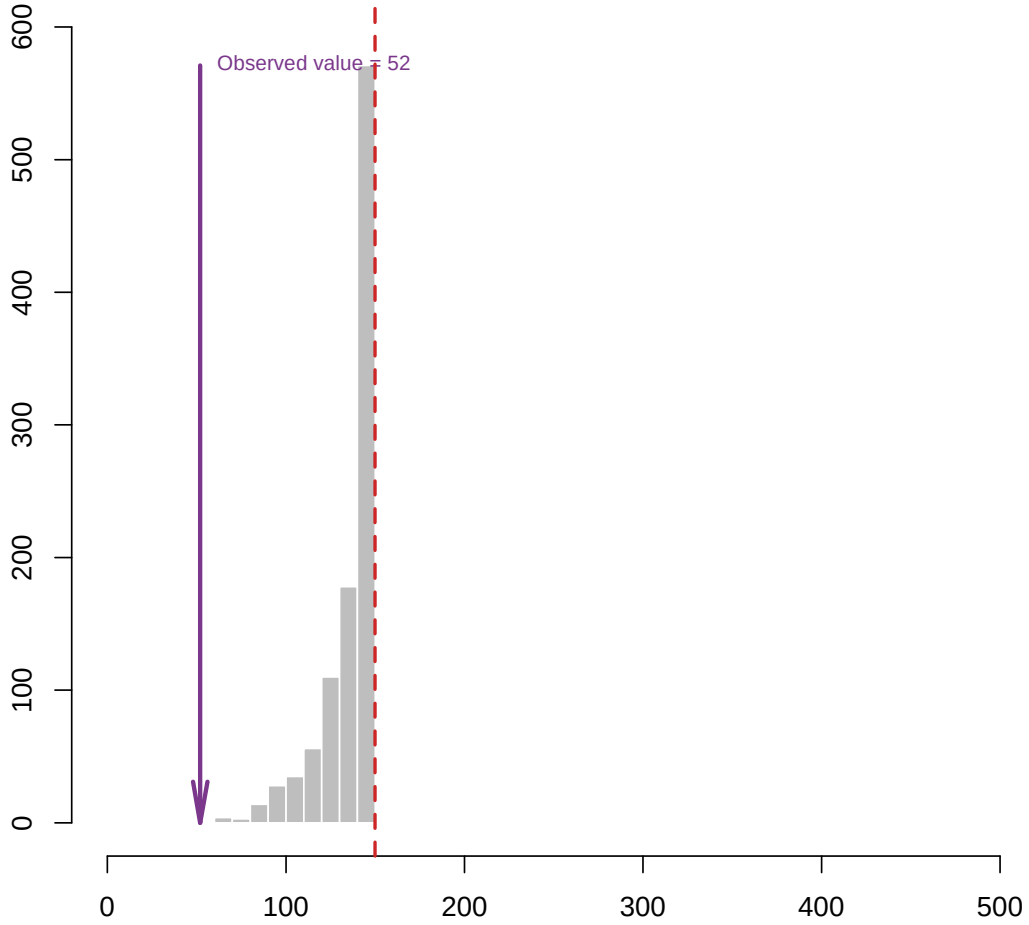
Repeat all that a lot (1,000 times)

Get the distribution

# First Component (explains 392 out of 444). $P < .001$



# Second Component (explains 52 out of 444). Very *n.s.*



# How Many Components?

Scree Test

Eigen-Values Larger than 1

Eigen-Values Larger than average

Parallel Test

Etc.

# Bootstrap

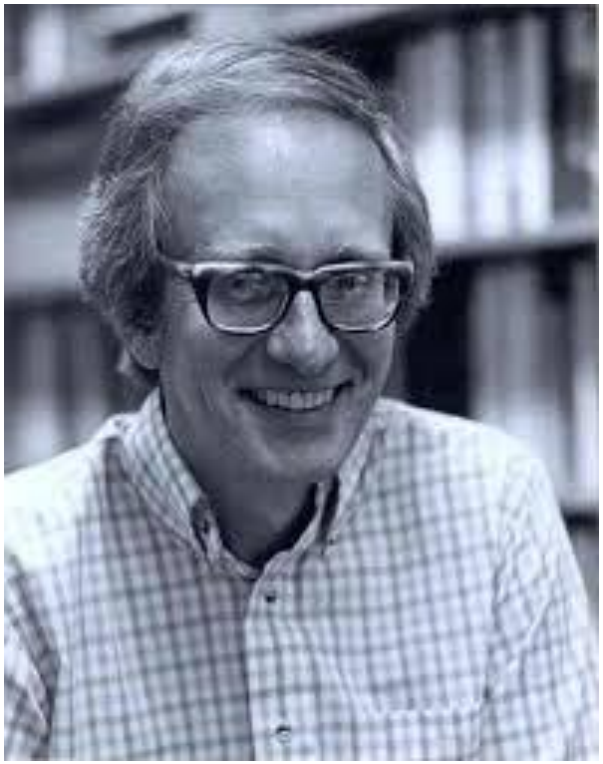
Important variables for a component

Create new samples, and see what stays

This is called the "Bootstrap"

# Bootstrap Test

**Bradley Efron**



# Standard Model

- *Ideally* (Population):
  - Look for “large loading” variables
- How to find them:
  - Sample repeatedly from this *infinite* Population
  - Infinity big property: probabilities do *not* change

- *No Population*:
  - What to do then?
- Replace *population* by *sample*
  - Sample repeatedly from *finite* sample
  - Make sure that probabilities do not change
    - This means

**sample *with* replacement!**

# Bootstrap sampling: Take the 20 wines

# Bootstrap sampling:

Take the 20 wines.

I mean the numbers 1 to 20 …

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

# And the Bootstrap Hat

**Sampling with Replacement:**
***p* does not change!**

7 8 11 13 16
5 10
1 6 12 17
9 15 19
2 3 4 18
14 20

**Sampling with Replacement:**

***p* does not change!**

7 8 11 13 16
5 10
1 6 12 17
9 15 19
2 3 4 18
14 20

**Sampling with Replacement:**

***p* does not change!**


**Sampling with Replacement:**

***p* does not change!**

# A Shake!

10

10

10


10

# And so on …
# Till we have 20 Observations

10 3 10 12 19 12 1 3 18 12 17 6 18 1 10 9 5 3 19 6

## Sampling with Replacement

## Some are repeated, some are lost

# Do that again

Run a PCA on new sample

Repeat that a lot (1,000 times)

Compute loadings for variables

Get confidence intervals

Compute means and standard deviations

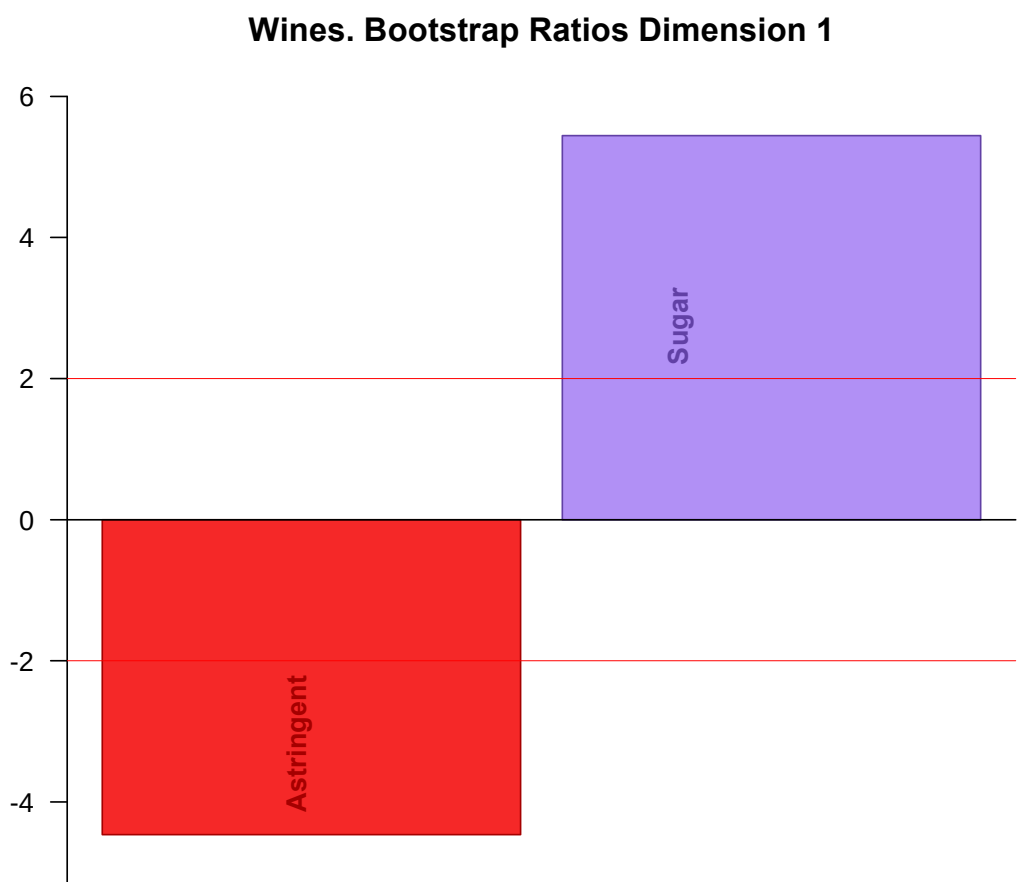
Bootrap ratios: $BR = M / S$

*BR* ARE LIKE STUDENT'S *T*

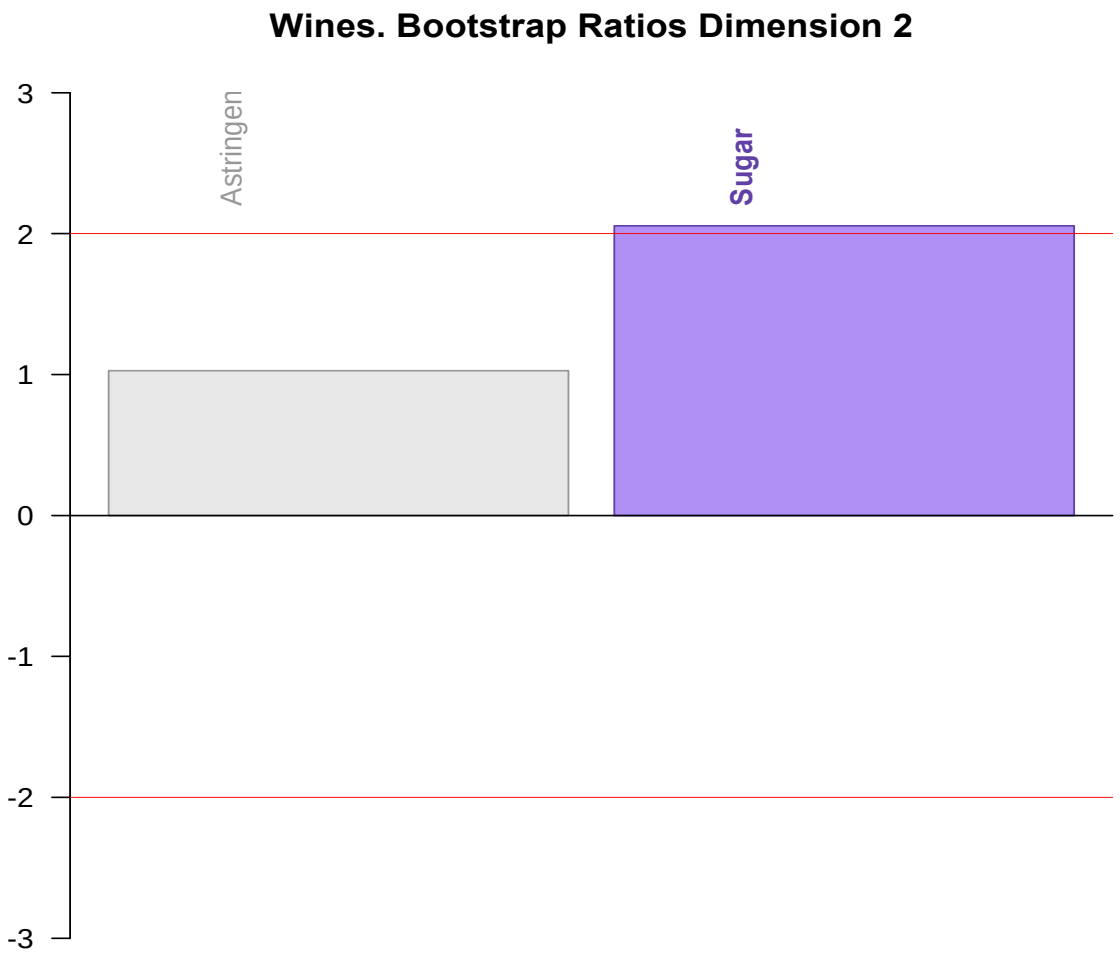
SO *BR* > 2 SAYS *P* < .05

... AND, OF COURSE, THE EARTH IS ROUND!

# Bootstrap ratios. PC1. BR > 2, *P* < .05



# BOOTSTRAP RATIOS. PC2. BR > 2, *P* < .05



# The story so far

What have learn so far?

We can describe multivariate data (with pictures)

Inference 1:
What is real (permutation)

Inference 2:
What is important (Bootstrap)

# This is the end, my friend!...

Thank you all!

Merci …

# Your Turn

Questions?

Comments?

# Additional Slides:

What is behind
the eigen-fairy,
the SVD fairy
and
the magic Lagrangian …

## A Maximization problem

$\mathbf{X}$ is an $I$ by $J$ matrix

We want vectors:

$I \times 1$ $\mathbf{p}$ and $J \times 1$ $\mathbf{q}$

such that:

$$\arg\max_{\mathbf{pq}} \delta = \mathbf{p}^\top \mathbf{X} \mathbf{q} \text{ with } \mathbf{p}^\top \mathbf{p} = \mathbf{q}^\top \mathbf{q} = 1$$

Equivalent to

$$\mathbf{X}\mathbf{q} = \delta\mathbf{p}$$

$$\mathbf{X}^\top\mathbf{p} = \delta\mathbf{q}.$$

...

$$\mathbf{X}\mathbf{q} = \delta\mathbf{p}$$

$$\mathbf{X}^\top\mathbf{p} = \delta\mathbf{q}.$$

and also equivalent to

$$\mathbf{q}^\top\mathbf{X}^\top\mathbf{X}\mathbf{q} = \delta^2\mathbf{p}^\top\mathbf{p} = \delta^2$$

$$\mathbf{p}^\top\mathbf{X}\mathbf{X}^\top\mathbf{p} = \delta^2\mathbf{q}^\top\mathbf{q} = \delta^2.$$

## The LAGRANGIAN!

$$\mathcal{L} = \mathbf{q}^\top \mathbf{X}^\top \mathbf{X}\mathbf{q} + \mathbf{p}^\top \mathbf{X}\mathbf{X}^\top \mathbf{p} - 2\delta^2 - \alpha\left(\mathbf{q}^\top \mathbf{q} - 1\right) - \beta\left(\mathbf{p}^\top \mathbf{p} - 1\right)$$

Partial derivatives of $\mathcal{L}$ for $\mathbf{q}$ and $\mathbf{p}$

Partial derivatives of $\mathcal{L}$ for $\mathbf{q}$ and $\mathbf{p}$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{q}} = 2\mathbf{X}^\top \mathbf{X}\mathbf{q} - 2\alpha\mathbf{q}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{p}} = 2\mathbf{X}\mathbf{X}^\top \mathbf{p} - 2\beta\mathbf{p} \ .$$

## Set the Partial Derivatives to 0

$$\mathbf{X}^\top \mathbf{X}\mathbf{q} = \alpha \mathbf{q}$$

$$\mathbf{X}\mathbf{X}^\top \mathbf{p} = \beta \mathbf{p} \ .$$

## Looks like an Eigen-Problem!

# Eigen?

$$\mathbf{X}^\top \mathbf{X}\mathbf{q} = \alpha \mathbf{q} \qquad \mathbf{X}\mathbf{X}^\top \mathbf{p} = \beta \mathbf{p}$$