

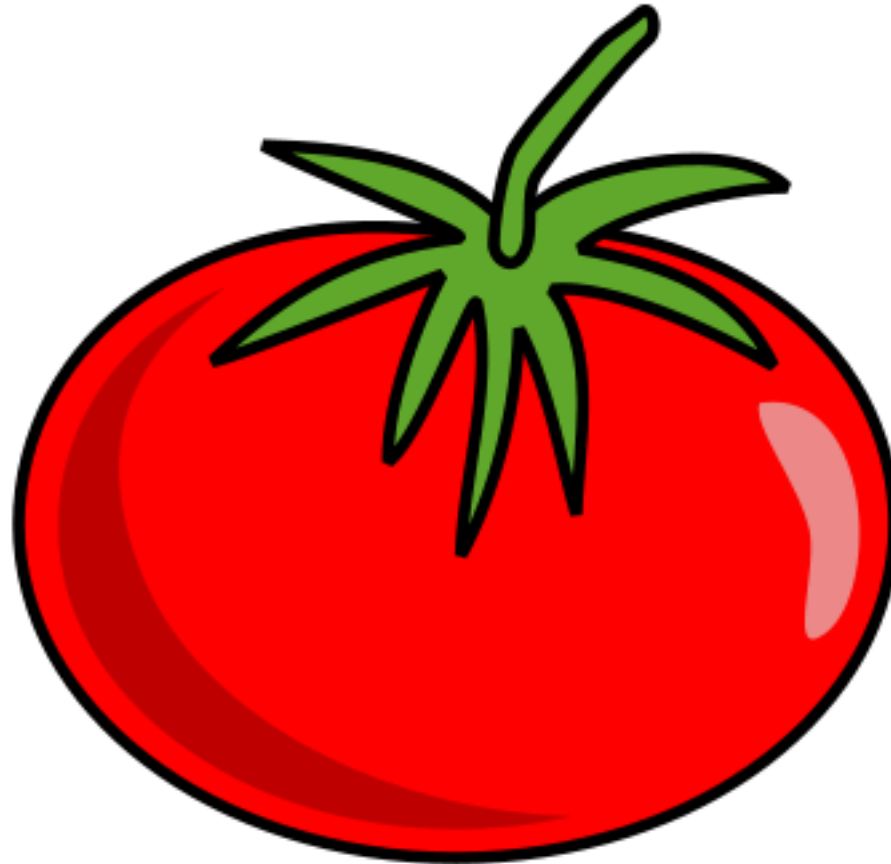
# Principle Component Analysis

Sébastien Lê

# Principle Component Analysis

- What does the name of this method remind you of?
- Which kinds of applications do you imagine?

# Example



# Example

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	8,43	5,46	4,50	0,50	6,89	8,00
D	4,79	2,39	7,07	6,39	3,96	4,29
F	6,79	6,96	3,29	1,61	4,11	5,46
G	4,25	6,39	3,96	5,07	2,89	3,93
H	6,89	6,32	3,61	2,46	4,18	6,25
O	8,32	4,93	4,79	1,00	6,46	7,29
Q	5,71	2,36	7,64	2,14	3,61	4,43

# Example

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	8,43	5,46	4,50	0,50	6,89	8,00
D	4,79	2,39	7,07	6,39	3,96	4,29
F	6,79	6,96	3,29	1,61	4,11	5,46
G	4,25	6,39	3,96	5,07	2,89	3,93
H	6,89	6,32	3,61	2,46	4,18	6,25
O	8,32	4,93	4,79	1,00	6,46	7,29
Q	5,71	2,36	7,64	2,14	3,61	4,43
Mean	6,45	4,97	4,98	2,74	4,59	5,66
Sdt. Dev.	1,51	1,75	1,58	2,02	1,39	1,46

# Example

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	8,43	5,46	4,50	0,50	6,89	8,00
D	4,79	2,39	7,07	6,39	3,96	4,29
F	6,79	6,96	3,29	1,61	4,11	5,46
G	4,25	6,39	3,96	5,07	2,89	3,93
H	6,89	6,32	3,61	2,46	4,18	6,25
O	8,32	4,93	4,79	1,00	6,46	7,29
Q	5,71	2,36	7,64	2,14	3,61	4,43
Mean	6,45	4,97	4,98	2,74	4,59	5,66
Sdt. Dev.	1,51	1,75	1,58	2,02	1,39	1,46

Which one is the most **FIRM**?

Which one is the less **SWEET**?

Can I say that **F** is more **FIRM** than **G** is less **SWEET**?  
(does it make sense?)

# When data are mean-centered

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	1,97	0,49	-0,48	-2,24	2,31	2,34
D	-1,67	-2,58	2,09	3,65	-0,62	-1,38
F	0,33	1,99	-1,69	-1,13	-0,48	-0,20
G	-2,20	1,42	-1,02	2,33	-1,69	-1,73
H	0,44	1,35	-1,37	-0,28	-0,41	0,59
O	1,87	-0,05	-0,19	-1,74	1,88	1,62
Q	-0,74	-2,62	2,66	-0,60	-0,98	-1,23

Which one is the most **FIRM**?

Which one is the less **SWEET**?

Can I say that **F** is more **FIRM** than **G** is less **SWEET**?  
(does it make sense?)

# When data are scaled to unit variance

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	1,31	0,28	-0,30	-1,11	1,66	1,60
D	-1,11	-1,47	1,32	1,81	-0,45	-0,94
F	0,22	1,13	-1,07	-0,56	-0,34	-0,14
G	-1,46	0,81	-0,64	1,15	-1,22	-1,19
H	0,29	0,77	-0,87	-0,14	-0,29	0,40
O	1,24	-0,03	-0,12	-0,86	1,35	1,11
Q	-0,49	-1,49	1,68	-0,30	-0,71	-0,85

Which one is the most **FIRM**?

Which one is the less **SWEET**?

Can I say that **F** is more **FIRM** than **G** is less **SWEET**?  
(does it make sense?)

# When data are scaled to unit variance

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	1,31	0,28	-0,30	-1,11	1,66	1,60
D	-1,11	-1,47	1,32	1,81	-0,45	-0,94
F	0,22	1,13	-1,07	-0,56	-0,34	-0,14
G	-1,46	0,81	-0,64	1,15	-1,22	-1,19
H	0,29	0,77	-0,87	-0,14	-0,29	0,40
O	1,24	-0,03	-0,12	-0,86	1,35	1,11
Q	-0,49	-1,49	1,68	-0,30	-0,71	-0,85

What do you think about A and O?

# When data are scaled to unit variance

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	1,31	0,28	-0,30	-1,11	1,66	1,60
D	-1,11	-1,47	1,32	1,81	-0,45	-0,94
F	0,22	1,13	-1,07	-0,56	-0,34	-0,14
G	-1,46	0,81	-0,64	1,15	-1,22	-1,19
H	0,29	0,77	-0,87	-0,14	-0,29	0,40
O	1,24	-0,03	-0,12	-0,86	1,35	1,11
Q	-0,49	-1,49	1,68	-0,30	-0,71	-0,85

What do you think about **F** and **H**?

# When data are scaled to unit variance

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	1,31	0,28	-0,30	-1,11	1,66	1,60
D	-1,11	-1,47	1,32	1,81	-0,45	-0,94
F	0,22	1,13	-1,07	-0,56	-0,34	-0,14
G	-1,46	0,81	-0,64	1,15	-1,22	-1,19
H	0,29	0,77	-0,87	-0,14	-0,29	0,40
O	1,24	-0,03	-0,12	-0,86	1,35	1,11
Q	-0,49	-1,49	1,68	-0,30	-0,71	-0,85

What do you think about A and D?

# When data are scaled to unit variance

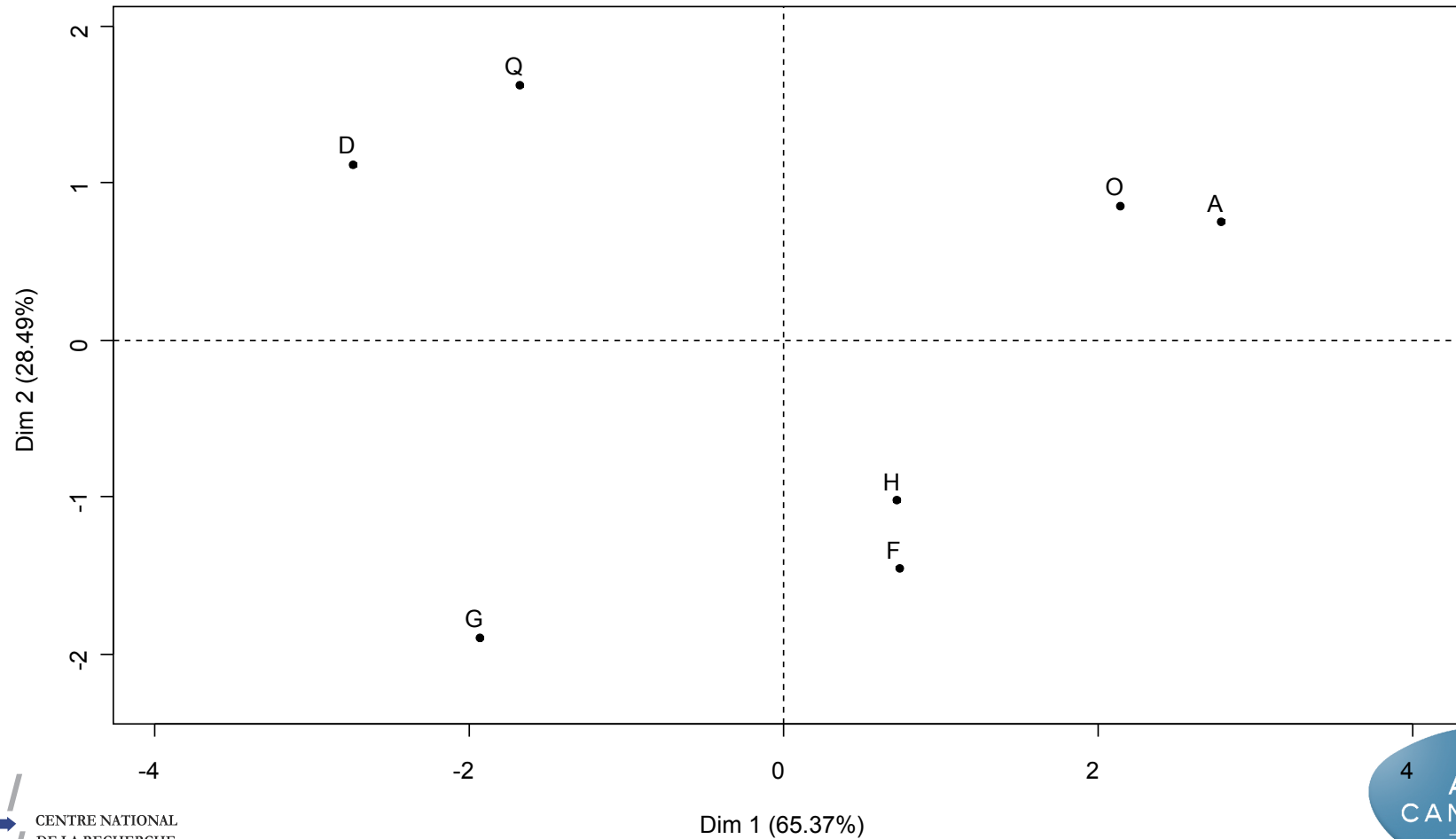
Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	1,31	0,28	-0,30	-1,11	1,66	1,60
D	-1,11	-1,47	1,32	1,81	-0,45	-0,94
F	0,22	1,13	-1,07	-0,56	-0,34	-0,14
G	-1,46	0,81	-0,64	1,15	-1,22	-1,19
H	0,29	0,77	-0,87	-0,14	-0,29	0,40
O	1,24	-0,03	-0,12	-0,86	1,35	1,11
Q	-0,49	-1,49	1,68	-0,30	-0,71	-0,85

What do you think about Q and F?

# Example

# Example

Individuals factor map (PCA)



# When data are scaled to unit variance

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor	
A	1,31		0,28	-0,30	-1,11	1,66	1,60
D	-1,11		-1,47	1,32	1,81	-0,45	-0,94
F	0,22		1,13	-1,07	-0,56	-0,34	-0,14
G	-1,46		0,81	-0,64	1,15	-1,22	-1,19
H	0,29		0,77	-0,87	-0,14	-0,29	0,40
O	1,24		-0,03	-0,12	-0,86	1,35	1,11
Q	-0,49		-1,49	1,68	-0,30	-0,71	-0,85

What do you think about **EXT\_COL**, **SWEET** and **TOMATO\_FLAVOR**?

# When data are scaled to unit variance

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	1,31	0,28	-0,30	-1,11	1,66	1,60
D	-1,11	-1,47	1,32	1,81	-0,45	-0,94
F	0,22	1,13	-1,07	-0,56	-0,34	-0,14
G	-1,46	0,81	-0,64	1,15	-1,22	-1,19
H	0,29	0,77	-0,87	-0,14	-0,29	0,40
O	1,24	-0,03	-0,12	-0,86	1,35	1,11
Q	-0,49	-1,49	1,68	-0,30	-0,71	-0,85

What do you think about **EXT\_COLOR** and **MEALY**?

# When data are scaled to unit variance

Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	1,31	0,28	-0,30	-1,11	1,66	1,60
D	-1,11	-1,47	1,32	1,81	-0,45	-0,94
F	0,22	1,13	-1,07	-0,56	-0,34	-0,14
G	-1,46	0,81	-0,64	1,15	-1,22	-1,19
H	0,29	0,77	-0,87	-0,14	-0,29	0,40
O	1,24	-0,03	-0,12	-0,86	1,35	1,11
Q	-0,49	-1,49	1,68	-0,30	-0,71	-0,85

What do you think about **FIRM** and **MELTY**?

# When data are scaled to unit variance

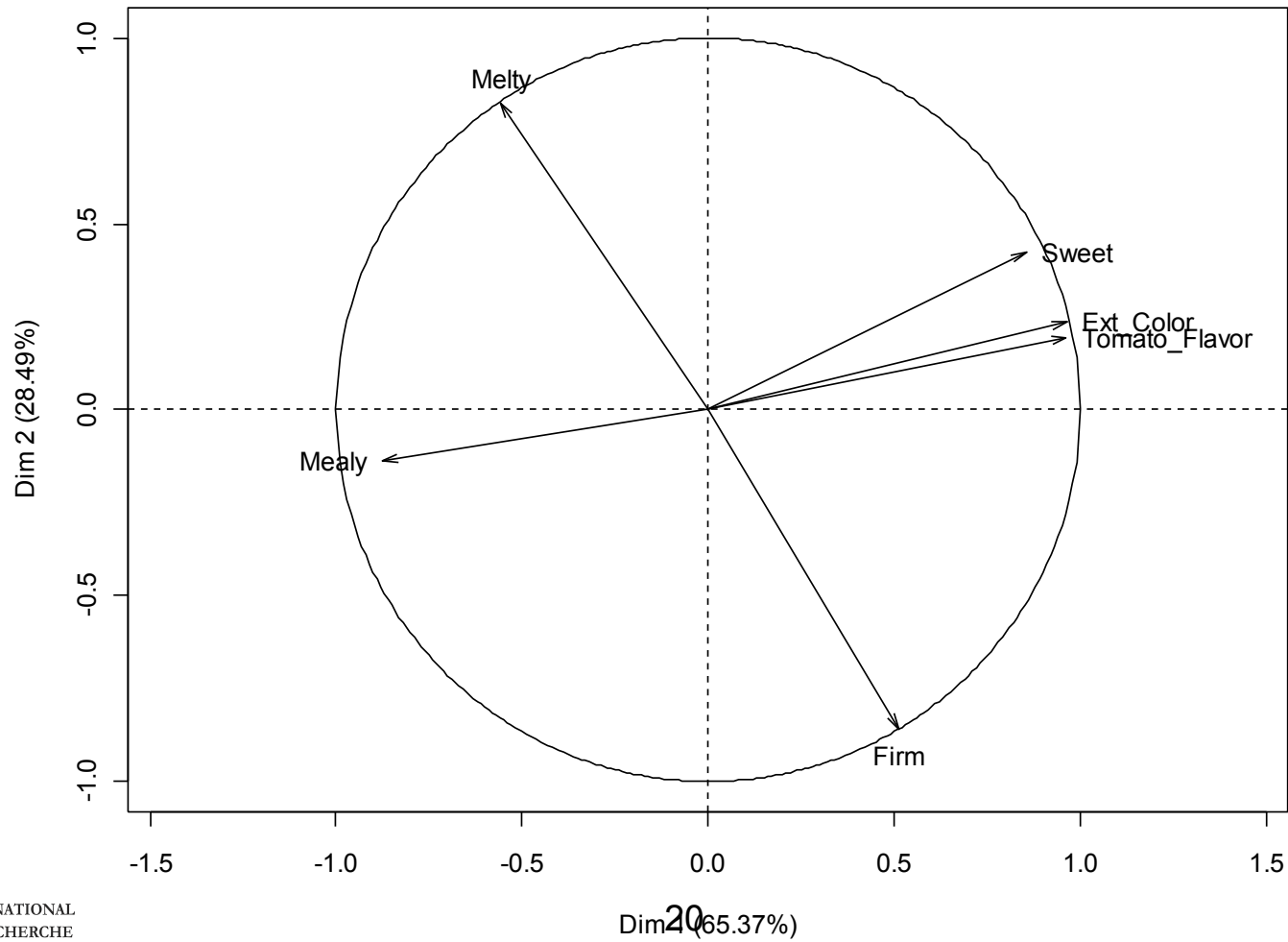
Tomatoes	Ext_Color	Firm	Melty	Mealy	Sweet	Tomato_Flavor
A	1,31	0,28	-0,30	-1,11	1,66	1,60
D	-1,11	-1,47	1,32	1,81	-0,45	-0,94
F	0,22	1,13	-1,07	-0,56	-0,34	-0,14
G	-1,46	0,81	-0,64	1,15	-1,22	-1,19
H	0,29	0,77	-0,87	-0,14	-0,29	0,40
O	1,24	-0,03	-0,12	-0,86	1,35	1,11
Q	-0,49	-1,49	1,68	-0,30	-0,71	-0,85

What do you think about SWEET and MELTY?

# Example

# Example

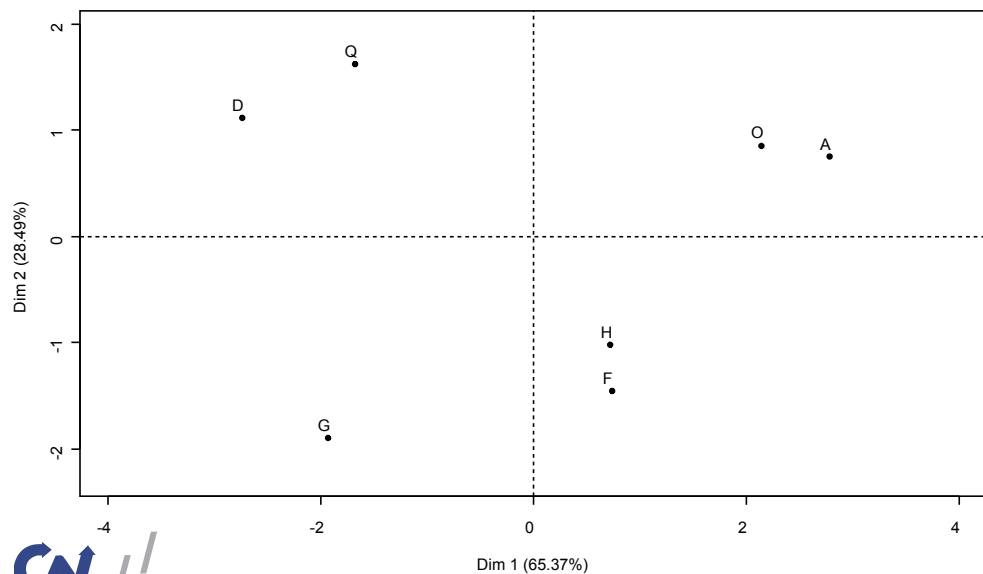
Variables factor map (PCA)



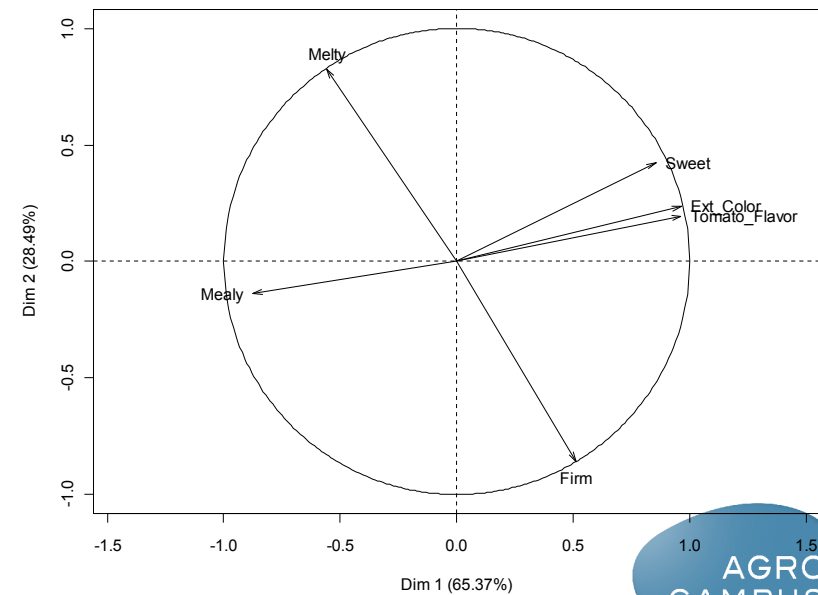
# Congratulations!

- You've just made your first PCA by hand
- With the scatter plot of the individuals (tomatoes)
- With the scatter plot of the variables (sensory descriptors)

Individuals factor map (PCA)

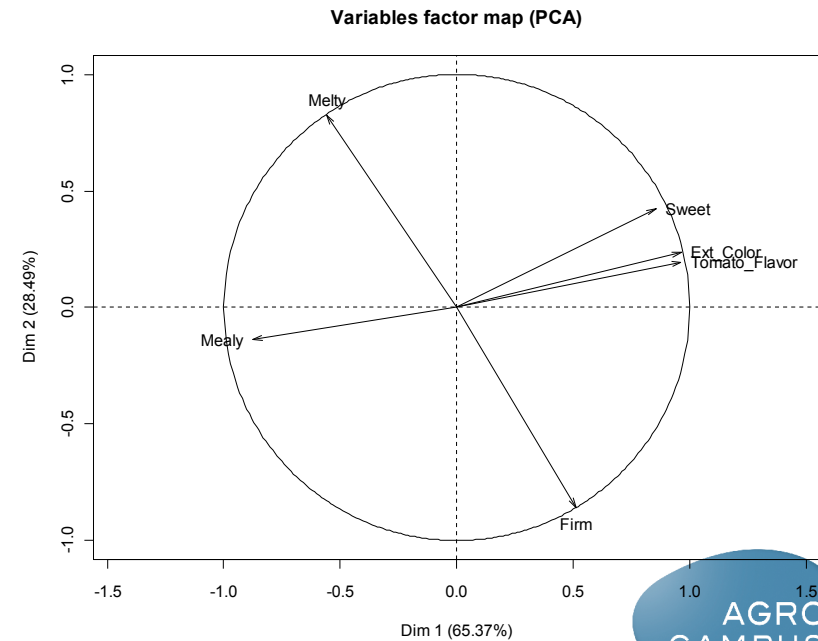
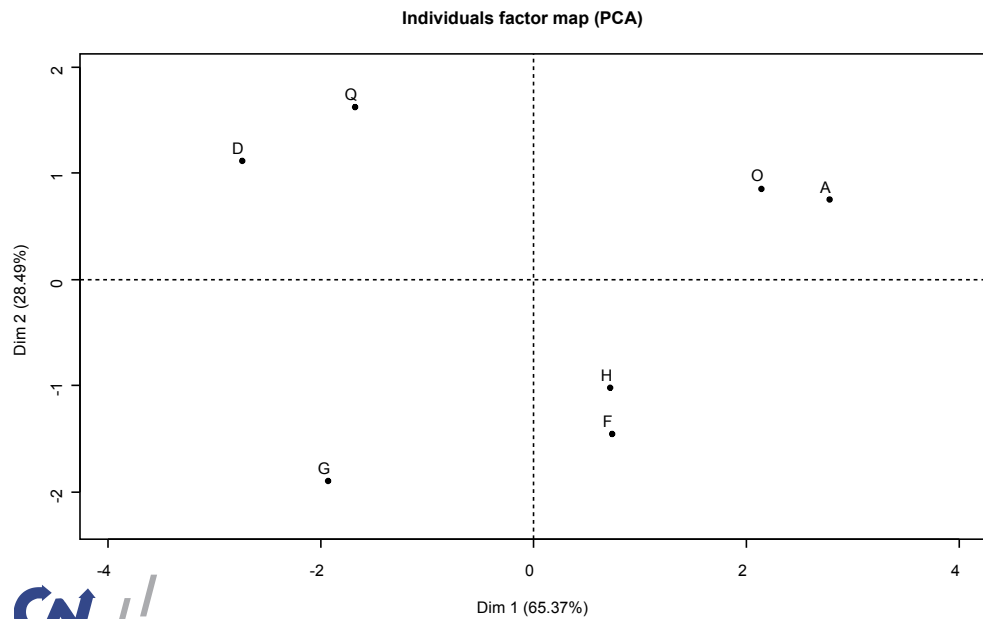


Variables factor map (PCA)



# Congratulations!

- Both outputs have to be interpreted jointly
- They summarize the same amount of information



# What is PCA?

- A statistical technique used to transform a number of **correlated** variables into a smaller number of “**uncorrelated**” variables called *principal components*
- The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible

This presentation is licensed under a  
[Creative Commons Attribution 4.0  
International License.](https://creativecommons.org/licenses/by/4.0/)



By Sébastien Lê