Data science in interdisciplinary research projects

Sébastien Déjean

Research engineer in statistics

math.univ-toulouse.fr/~sdejean

Masterclass MINT, December, 13th, 2023

Mathematics and Interactions in Toulouse

INSTITUT de MATHEMATIQUES de TOULOUSE

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Université de Toulouse

Who am I? Short resume

1993
- A levels, Mathematics and Physics

1996
- Bachelor of Science, Industrial Mathematics

1998
- Master degree, Applied Mathematics

2002
- PhD in Applied Mathematics, Statistics

Comparison of estimation methods for nonlinear mixed effects models. Application to the modelling of the evolution of the leaf area index of crops observed by remote sensing

2003
- Hired as a **research engineer in Statistics** at Université Toulouse III – Paul Sabatier

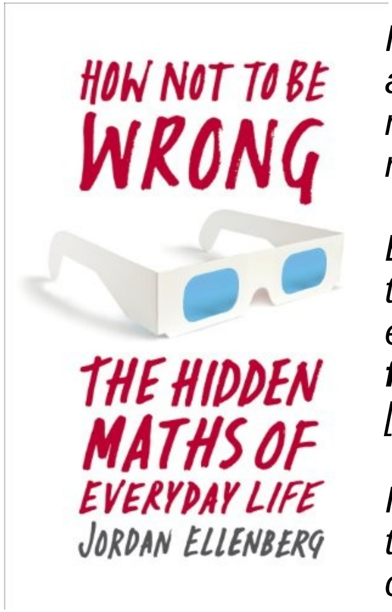
2019
- Habilitation Thesis

Fifteen years of applied research in data science

What does a research engineer in statistics do?

- Officially (in French, `data.enseignementsup-recherche.gouv.fr/pages/fiche_emploi_type_referens_iii_itrf/?refine.referens_id=E1D44`)
  - engineer responsible for statistical aspects in a research laboratory
  - manage statistical projects
  - define a data collection and management plan and the associated processing chain
  - participate in national and international research projects and associated publications
  - ...
- In practice
  - answer questions in the real world

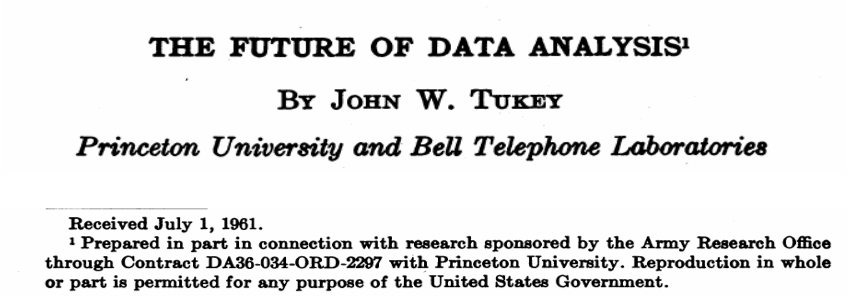
Address real-world problems



I blame word problems. They give a badly wrong impression of the relation between mathematics and reality. "Bobby has three hundred marbles and gives 30% of them to Jenny. He gives half as many to Jimmy as he gave to Jenny. How many does he have left?" That looks like it's about the real world, but it's just an arithmetic problem in a not very convincing disguise [...]

*But real-world questions aren't like word problems. A **real-world problem** is something like "Has the recession and its aftermath been especially bad for women in the workforce, and if so, to what extent is this the result of Obama administration policies?" **Your calculator doesn't have a button for this**. Because in order to give a sensible answer, you need to know more than just numbers. […]*

*It's only after you've started to formulate these questions that you take out the calculator. But at that point the real mental work is already finished. Dividing one number by another is mere computation; **figuring out what you should divide by what is mathematics.***



THE FUTURE OF DATA ANALYSIS[1]

BY JOHN W. TUKEY

Princeton University and Bell Telephone Laboratories

Received July 1, 1961.

[1] Prepared in part in connection with research sponsored by the Army Research Office through Contract DA36-034-ORD-2297 with Princeton University. Reproduction in whole or part is permitted for any purpose of the United States Government.

*The future of data analysis can involve great process, the overcoming of real difficulties, and the provision of a great service to all fields of science and technology. Will it? That remains to us, to our willingness to take up the **the rocky road of real problems** in preference to smooth road of unreal assumptions, arbitrary criteria, and abstract results without real attachments.*

The rocky road of real problems...

… smooth road of unreal assumptions, arbitrary criteria, and abstract results without real attachments

A real-world problem

How is global warming affecting plant growth?

Back to linear algebra

*The biological sciences are **today** in the process of changing from being primarily descriptive **to being very much quantitative**. As a result, biologists find themselves **confronted more and more with large amounts of numerical data** […]. But the mere collecting and recording of data achieve nothing; having been collected, they must be **investigated to see what information may be contained concerning the biological problem** at hand.[…]*

*Frequently, however, biologists have to subject their data to more complex calculations, requiring procedures that **involve mathematical details beyond their general experience**. In order to carry out the mathematics the biologist in this situation must either **learn the procedures himself**, or at least **learn something of the language of mathematics**, that he may **communicate satisfactorily with the mathematician** whose aid he enlists.*

S.R Searle (**1966**)
Matrix Algebra for the biological sciences

Eigen decomposition

$$M = V\Lambda V'$$

M a square matrix $n \times n$
V columns of V are the eigen vectors of M corresponding to the eigen values
Λ diagonal matrix of the eigen values $\boldsymbol{\lambda}_i$

$\boldsymbol{\lambda}$ is an eigen value of **M**
$\Leftrightarrow$
$\exists$ one vector **v** (length n) such that $\mathbf{Mv} = \boldsymbol{\lambda}\mathbf{v}$

Singular Value Decomposition (SVD)

$$M = U\Sigma V'$$

M a rectangular matrix m x n
U, V columns of U and V are, respectively, left and right singular vectors corresponding to singular values
Σ diagonal matrix of the singular values of M

σ is a singular value of **M**
$\Leftrightarrow$
$\exists$ 2 vectors **u** (length **m**) and **v** (length **n**) such that **Mv = σu** and **M'u = σv**

Link between eigen and SVD decompositions

$M = U\Sigma V'$ (SVD)

Let's compute : $M'M$ and MM'

$$
\begin{aligned}
M'M &= (U\Sigma V')'U\Sigma V' && \text{replace M with SVD decomposition} \\
&= V\Sigma'U'U\Sigma V' && (AB)' = B'A' \\
&= V\Sigma'\Sigma V' && \text{U unit vector, } U'U=UU'=I
\end{aligned}
$$

$$MM' = U\Sigma V'(U\Sigma V')' = U\Sigma V'V\Sigma'U' = U\Sigma\Sigma'U'$$

We obtain the eigen decomposition of M'M and MM' with eigen values equal to the square of the singular values and eigen vectors respectively equal to left and right singular vectors.

Linear algebra for statistics

Principal Component Analysis (PCA)

- X a nxp data matrix
- PCA is an orthogonal linear transformation that projects the data in a new coordinate system such that the greatest variance of the data lies on the first coordinate (first PC), the second greatest variance on the second PC and so on…
- It can be shown that :
  - ➔ The greatest variance is the first eigen value of X’X
  - ➔ Transforming coordinates is done using the first eigen vector

Jolliffe, I. T. (1986). Principal Component Analysis. Springer Series in Statistics. Springer-Verlag. doi:10.1007/b98835 ISBN 978-0-387-95442-4.

Linear algebra for statistics

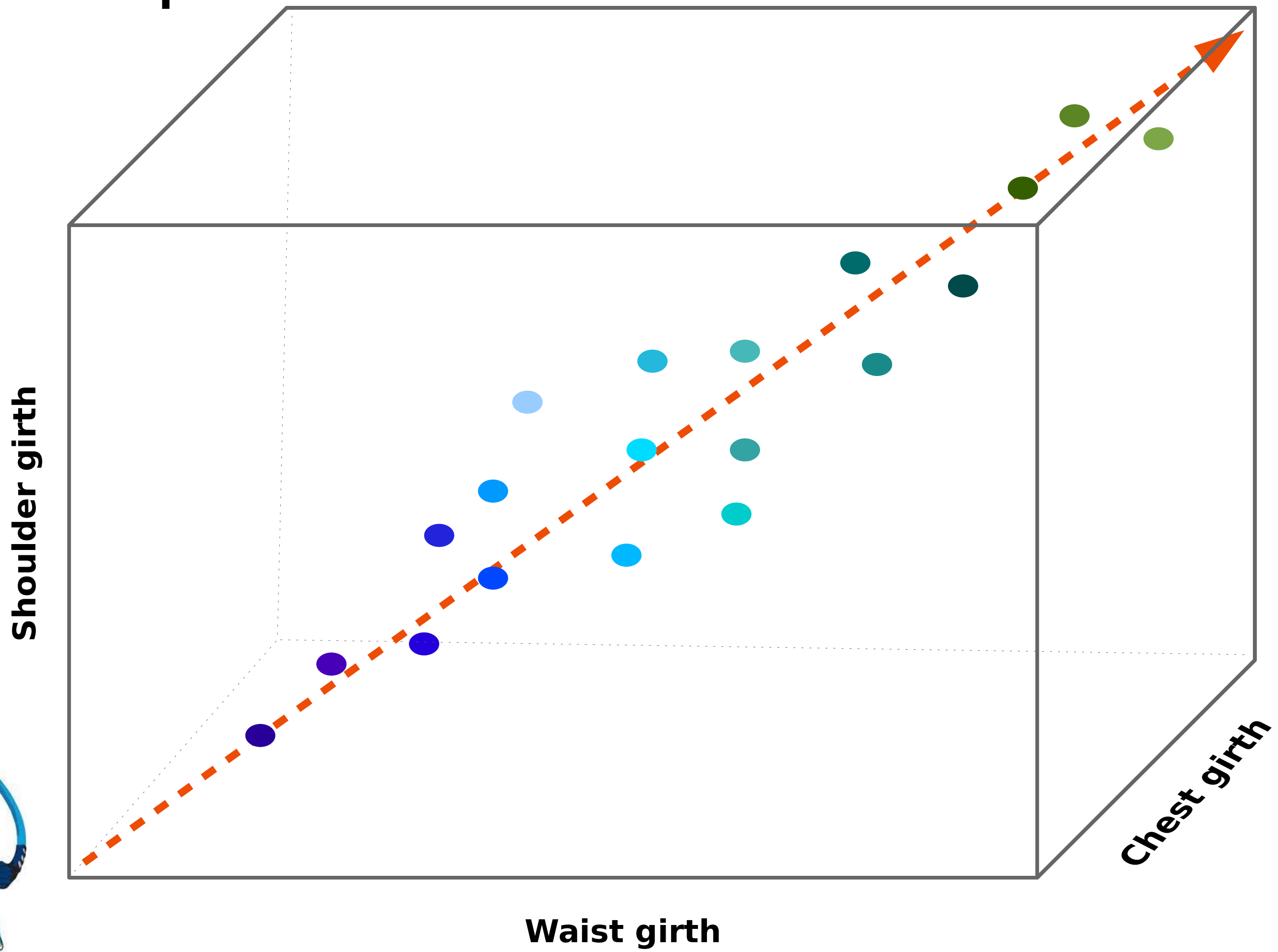
The **singular values** (in Σ) are the **square roots of the eigenvalues** of the matrix **X'X**. Each **eigenvalue is proportional to the portion of the "variance"** (more correctly of the sum of the squared distances of the points from their multidimensional mean) that is associated with each eigenvector. The sum of all the eigenvalues is equal to the sum of the squared distances of the points from their multidimensional mean. **PCA essentially rotates the set of points around their mean in order to align with the principal components**. This **moves as much of the variance as possible** (using an orthogonal transformation) into the first few dimensions. The values in the remaining dimensions, therefore, tend to be small and may be dropped with minimal loss of information (see below). PCA is often used in this manner for **dimensionality reduction**. PCA has the distinction of being the optimal orthogonal transformation for keeping the subspace that has largest "variance" (as defined above).

en.wikipedia.org/wiki/Principal_component_analysis (2023/12/11)

PCA, principle

Teasing: Would you use a cubic box to pack a fishing rod?

PCA, principle



Do we need 3 dimensions to represent ‘standard’ individuals?

=

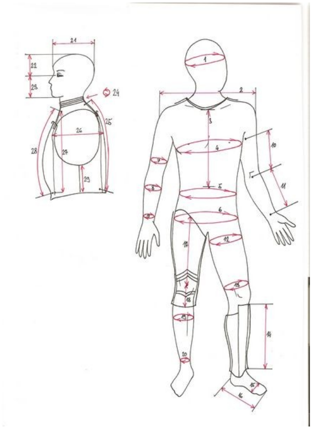
Do we need a cubic box to pack a fishing rod?

1st Principal Component: «beefyness»

PCA, toy example

- 20 individuals or observations
- 5 variables

```
s.g : shoulder girth (cm)
c.g : chest girth (cm)
w.g : waist girth (cm)
w   : weight (kg)
h   : height (cm)
```



Id	s.g	c.g	w.g	w	h
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	115.1	97.5	83.2	80.7	193.5
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	123.5	106.9	82.0	86.4	184.0
I8	120.4	102.5	76.8	78.4	184.5
I9	111.0	91.0	68.5	62.0	175.0
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	104.0	95.4	86.0	76.8	157.5
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	91.9	78.1	57.9	48.6	160.7
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8

PCA, toy example

Raw data

Id	s.g	c.g	w.g	w	h
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	115.1	97.5	83.2	80.7	193.5
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	123.5	106.9	82.0	86.4	184.0
I8	120.4	102.5	76.8	78.4	184.5
I9	111.0	91.0	68.5	62.0	175.0
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	104.0	95.4	86.0	76.8	157.5
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	91.9	78.1	57.9	48.6	160.7
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8

Covariance matrix

	s.g	c.g	w.g	w	h
s.g	**68.6**	37.7	28.1	55.3	61.2
c.g	37.7	**37.5**	33.9	45.7	32.4
w.g	28.1	33.9	**50.8**	56.6	27.7
w	55.3	45.7	56.6	**85.7**	59.5
h	61.2	32.4	27.7	59.5	**109.3**

$$68.6 + 37.5 + 50.8 + 85.7 + 109.3 = \mathbf{351.9}$$

351.9 represents the quantity of information contained in the data.

Eigen decomposition of the covariance matrix

www.r-project.org

```
R> eigen(cov(dataBody))
eigen() decomposition
$values
[1] 255.7  60.2  23.5   8.6   4.0

$vectors
      [,1]  [,2]  [,3]  [,4]  [,5]
[1,]  -0.45  0.16  0.78  0.18 -0.36
[2,]  -0.32 -0.25  0.26 -0.72  0.49
[3,]  -0.34 -0.53 -0.33 -0.24 -0.66
[4,]  -0.53 -0.36 -0.18  0.60  0.44
[5,]  -0.54  0.70 -0.42 -0.16 -0.02
```

```
R> prcomp(dataBody)

Standard deviations (1, .., p=5):
[1] 15.99  7.76  4.85  2.93  2.00

Rotation (n x k) = (5 x 5):
             PC1   PC2   PC3   PC4   PC5
shoulder.g 0.45 -0.16  0.78 -0.18  0.36
chest.g    0.32  0.25  0.26  0.72 -0.49
waist.g    0.34  0.53 -0.33  0.24  0.66
weight     0.54  0.36 -0.18 -0.60 -0.44
height     0.54 -0.71 -0.43  0.17  0.02
```

Coefficients of linear combinations or loadings

```
PC1 =  0.45*shoulder.g + 0.32*chest.g + 0.34*waist.g + 0.54*weight + 0.54*height
PC2 = -0.16*shoulder.g + 0.25*chest.g + 0.53*waist.g + 0.36*weight – 0.70*height
...
```

Transorm the data

Centered data

Ex: -6.50 = **0.45***(-1.9) + **0.32***(-4.7) + **0.34***(-3.8) + **0.54***(-5) + **0.54***(-0.4)

Id	s.g	c.g	w.g	w	h
I1	-1.9	-4.7	-3.8	-5.0	-0.4
I2	2.4	2.8	3.7	1.2	0.9
I3	7.0	3.3	7.9	10.1	19.1
I4	-3.6	2.8	2.5	2.0	12.1
I5	-0.6	3.3	4.7	8.2	12.8
I6	11.7	5.7	7.2	4.2	7.1
I7	15.4	12.7	6.7	15.8	9.6
I8	12.3	8.3	1.5	7.8	10.1
I9	2.9	-3.2	-6.8	-8.6	0.6
I10	11.4	-0.7	2.2	11.0	9.6
I11	-3.1	-5.2	-4.1	-3.3	-4.9
I12	-7.9	-0.1	4.2	4.9	-14.4
I13	-9.0	-3.4	2.6	-2.4	-1.7
I14	-0.5	2.8	-5.8	-9.2	-11.8
I15	-4.1	1.2	10.7	6.2	-16.9
I16	0.3	-2.4	-5.4	1.2	2.1
I17	-8.8	-6.9	-11.8	-15.1	-10.0
I18	-16.2	-16.1	-17.4	-22.0	-13.7
I19	-1.0	-3.3	-3.1	-4.2	-0.4
I20	-7.6	2.9	5.1	-3.3	-10.6

	PC1	PC2	PC3	PC4	PC5
s.g	0.45	-0.16	0.78	-0.18	0.36
c.g	0.32	0.25	0.26	0.72	-0.49
w.g	0.34	0.53	-0.33	0.24	0.66
w	0.54	0.36	-0.17	-0.60	-0.44
h	0.54	-0.70	-0.43	0.17	0.02

Apply loadings

	PC1	PC2	PC3	PC4	PC5
I1	-6.50	-4.48	-0.37	-1.03	1.27
I2	4.40	2.04	0.81	1.87	1.38
I3	22.66	-5.94	-6.18	0.11	1.97
I4	7.78	-5.24	-8.38	4.10	-1.74
I5	13.73	-2.67	-8.02	0.82	-2.15
I6	15.67	-0.15	4.49	2.33	4.40
I7	26.99	3.19	6.29	0.04	-3.08
I8	18.41	-3.43	5.63	1.09	-1.96
I9	-6.25	-8.48	4.97	0.79	1.86
I10	16.78	-3.67	1.99	-7.08	1.22
I11	-8.83	-0.78	0.28	-3.02	0.07
I12	-7.28	15.41	-2.31	-3.00	-2.35
I13	-6.45	2.25	-7.60	0.95	1.15
I14	-12.51	2.68	8.91	4.27	-1.53
I15	-3.65	20.76	-0.30	-2.45	1.99
I16	-0.63	-4.62	0.34	-3.46	-2.80
I17	-23.61	-5.07	2.20	1.19	-1.15
I18	-37.50	-9.07	-1.33	-1.89	-0.02
I19	-4.98	-3.61	0.33	-0.50	1.02
I20	-8.24	10.89	-1.74	4.86	0.44

Mean	**0**	**0**	**0**	**0**	**0**	
Var.	**255.7**	**60.2**	**23.5**	**8.6**	**4.0**	**= 351.9**

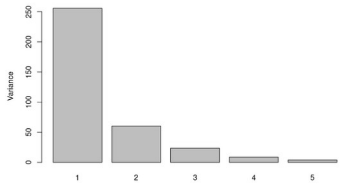
255.7 is the greatest variance we can obtain with a linear combination of the initial variables.

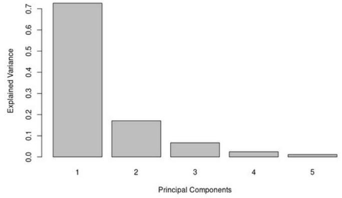
Graphical outputs (1/4)

	PC1	PC2	PC3	PC4	PC5
Variance	255.7	60.2	23.5	8.6	4.0
% variance	72.6	17.1	6.7	2.4	1.1

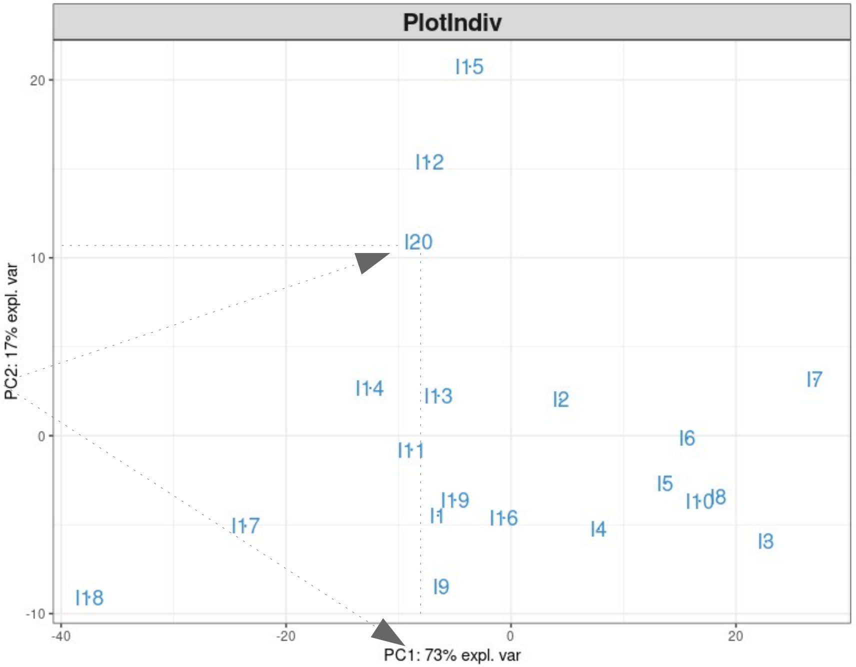
	PC1	PC2	PC3	PC4	PC5
I1	-6.50	-4.48	-0.37	-1.03	1.27
I2	4.40	2.04	0.81	1.87	1.38
I3	22.66	-5.94	-6.18	0.11	1.97
I4	7.78	-5.24	-8.38	4.10	-1.74
I5	13.73	-2.67	-8.02	0.82	-2.15
I6	15.67	-0.15	4.49	2.33	4.40
I7	26.99	3.19	6.29	0.04	-3.08
I8	18.41	-3.43	5.63	1.09	-1.96
I9	-6.25	-8.48	4.97	0.79	1.86
I10	16.78	-3.67	1.99	-7.08	1.22
I11	-8.83	-0.78	0.28	-3.02	0.07
I12	-7.28	15.41	-2.31	-3.00	-2.35
I13	-6.45	2.25	-7.60	0.95	1.15
I14	-12.51	2.68	8.91	4.27	-1.53
I15	-3.65	20.76	-0.30	-2.45	1.99
I16	-0.63	-4.62	0.34	-3.46	-2.80
I17	-23.61	-5.07	2.20	1.19	-1.15
I18	-37.50	-9.07	-1.33	-1.89	-0.02
I19	-4.98	-3.61	0.33	-0.50	1.02
I20	-8.24	10.89	-1.74	4.86	0.44

Screeplot





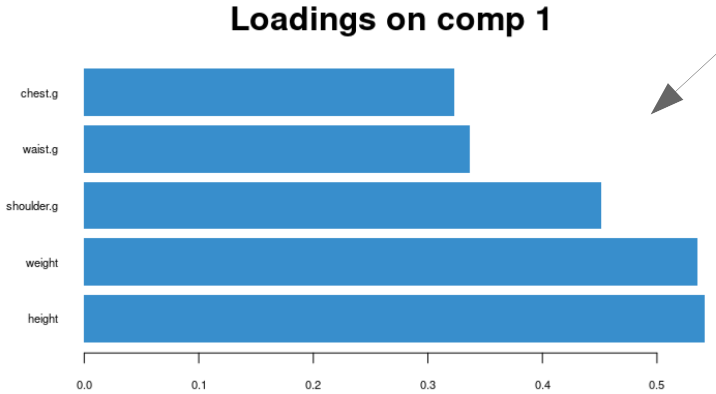
Individual plot



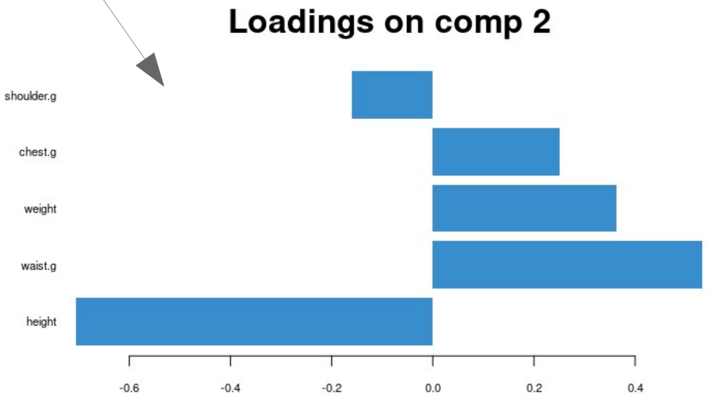
Graphical outputs (2/4)

Loadings

	PC1	PC2
shoulder.g	0.45	-0.16
chest.g	0.32	0.25
waist.g	0.34	0.53
weight	0.54	0.36
height	0.54	-0.70



Loading plot



Graphical outputs (3/4)

Id	s.g	c.g	w.g	w	h
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	115.1	97.5	83.2	80.7	193.5
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	123.5	106.9	82.0	86.4	184.0
I8	120.4	102.5	76.8	78.4	184.5
I9	111.0	91.0	68.5	62.0	175.0
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	104.0	95.4	86.0	76.8	157.5
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	91.9	78.1	57.9	48.6	160.7
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8

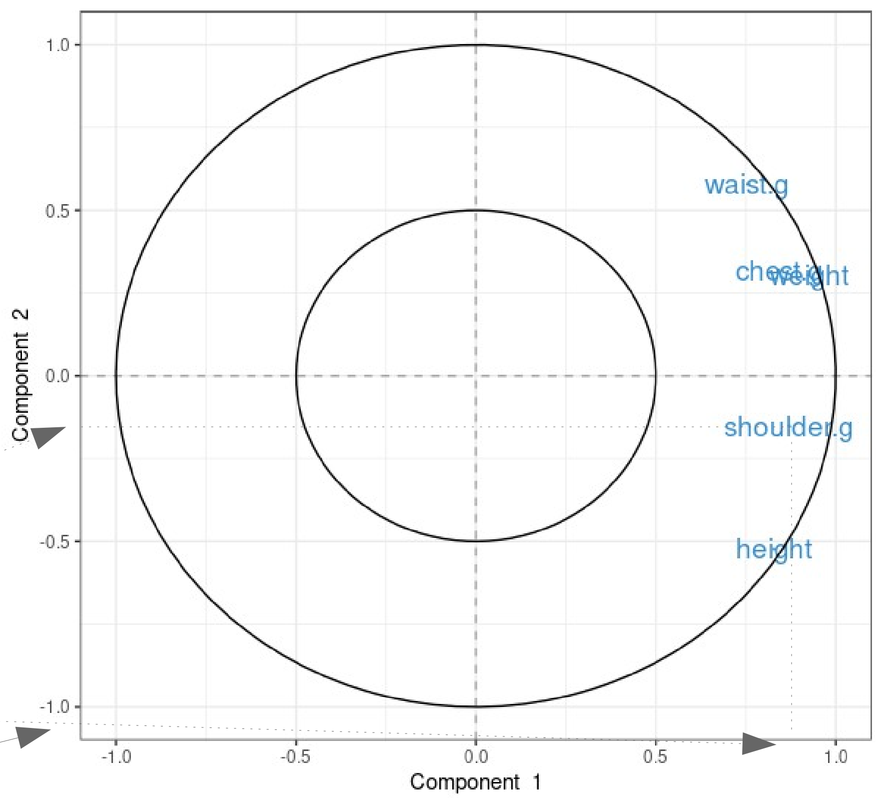
	PC1	PC2
I1	-6.50	-4.48
I2	4.40	2.04
I3	22.66	-5.94
I4	7.78	-5.24
I5	13.73	-2.67
I6	15.67	-0.15
I7	26.99	3.19
I8	18.41	-3.43
I9	-6.25	-8.48
I10	16.78	-3.67
I11	-8.83	-0.78
I12	-7.28	15.41
I13	-6.45	2.25
I14	-12.51	2.68
I15	-3.65	20.76
I16	-0.63	-4.62
I17	-23.61	-5.07
I18	-37.50	-9.07
I19	-4.98	-3.61
I20	-8.24	10.89

```
cor(s.g, PC1) =  0.87
cor(s.g, PC2) = -0,15

cor(c.g, PC1) =  0.84
cor(c.g, PC2) =  0.32
...
```

	PC1	PC2
shoulder.g	0.87	-0.15
chest.g	0.84	0.32
waist.g	0.75	0.58
weight	0.92	0.30
height	0.83	-0.52

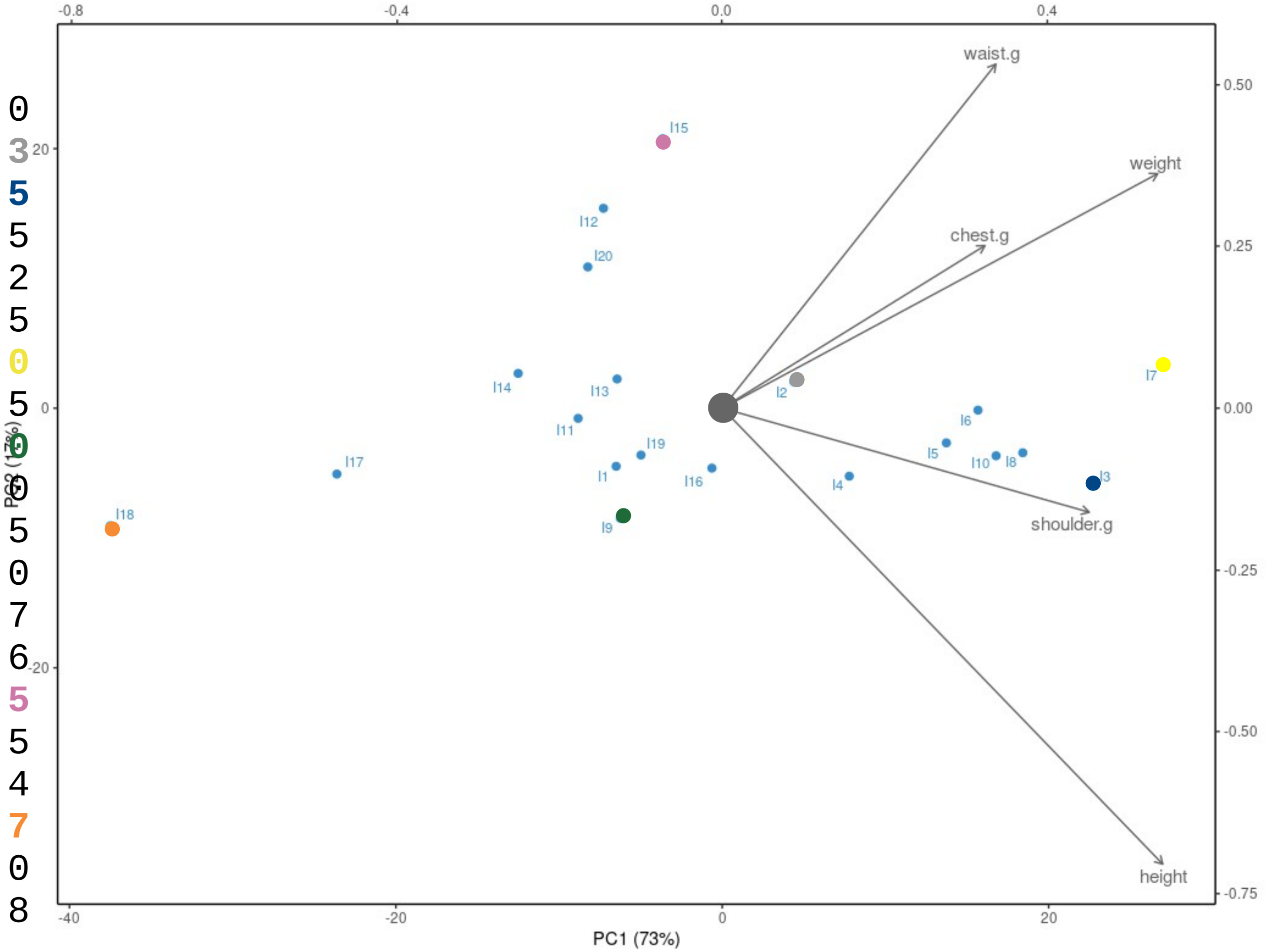
Correlation circle plot



Graphical outputs (4/4)

Id	s.g	c.g	w.g	w	h
I1	106.2	89.5	71.5	65.6	174.0
I2	110.5	97.0	79.0	71.8	175.3
I3	**115.1**	**97.5**	**83.2**	**80.7**	**193.5**
I4	104.5	97.0	77.8	72.6	186.5
I5	107.5	97.5	80.0	78.8	187.2
I6	119.8	99.9	82.5	74.8	181.5
I7	**123.5**	**106.9**	**82.0**	**86.4**	**184.0**
I8	120.4	102.5	76.8	78.4	184.5
I9	**111.0**	**91.0**	**68.5**	**62.0**	**175.0**
I10	119.5	93.5	77.5	81.6	184.0
I11	105.0	89.0	71.2	67.3	169.5
I12	100.2	94.1	79.6	75.5	160.0
I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6
I15	**104.0**	**95.4**	**86.0**	**76.8**	**157.5**
I16	108.4	91.8	69.9	71.8	176.5
I17	99.3	87.3	63.5	55.5	164.4
I18	**91.9**	**78.1**	**57.9**	**48.6**	**160.7**
I19	107.1	90.9	72.2	66.4	174.0
I20	100.5	97.1	80.4	67.3	163.8
Mean	108.1	94.2	75.3	70.6	174.4

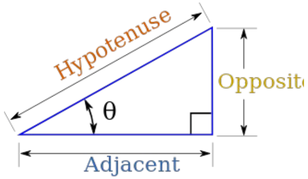
Biplot

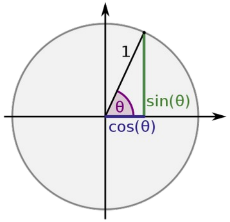


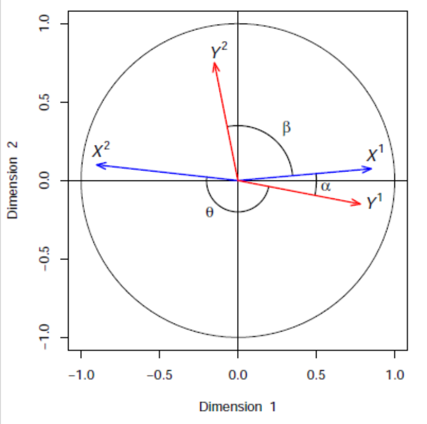
Focus on the variable plot

Correlation ↔ cosine

Remember trigonometry and right triangles:







The correlation between two variables is represented as:

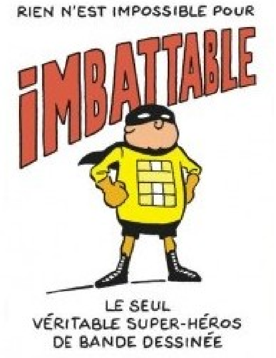
- An acute angle (cos(α) > 0) if it is positive
- An obtuse angle (cos(θ) < 0) if it is negative
- A right angle (cos(β)≈0) if it is near zero

Focus on the individual plot

- To interpret the graphical results of PCA must be done keeping in mind that one is looking at a projection on a plane (or in a volume for 3D representation)

- Be careful when interpreting visual proximities

- Illustration in comics with the only true super-heros ...

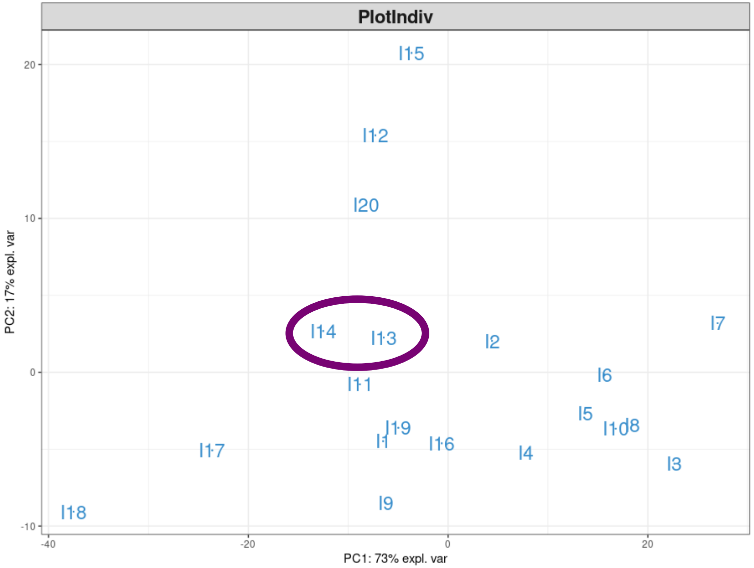


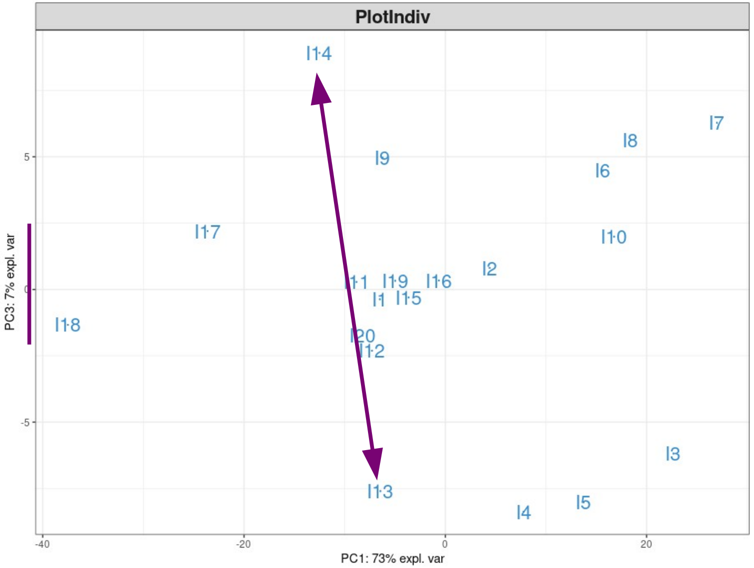
Scenario & illustration: Pascal Jousselin
Colour: Laurence Croix

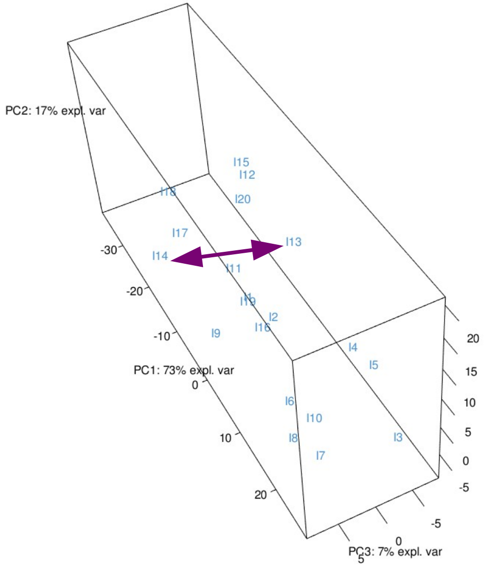
`pjousselin.free.fr`

Focus on the individual plot

I13	99.1	90.8	77.9	68.2	172.7
I14	107.6	97.0	69.6	61.4	162.6







Graphical outputs: summary

Screeplot

- How many components?
- 90% with 2 Pcs, 97% with 3PCs, 100% with 5PCs

Individual plot

- ‘Natural’ clusters, outliers...
- Caution: visual proximities

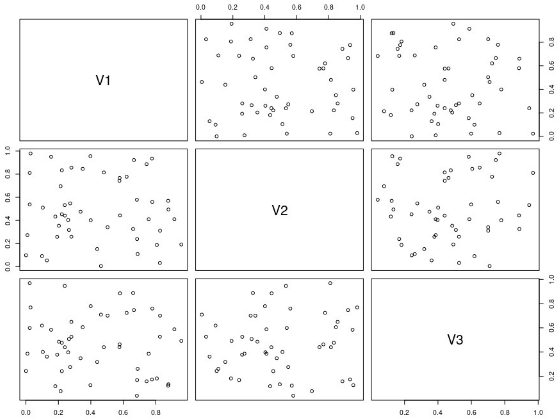
Variable plot, loading plot

- Correlation between variables
- Interpret components: PC1 « beefyness », PC2 « fatness, rotundity »

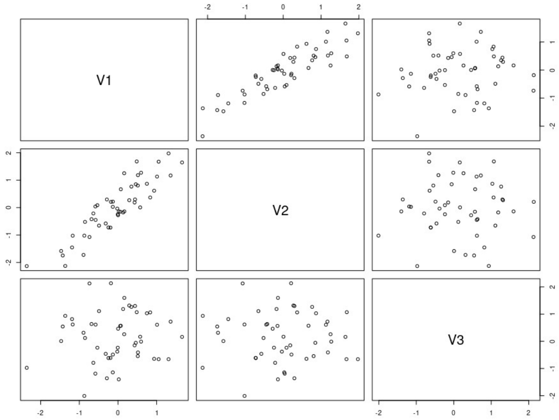
PCA, simulated examples

Data set : 50 observations, 3 variables (V1 – V2 - V3)

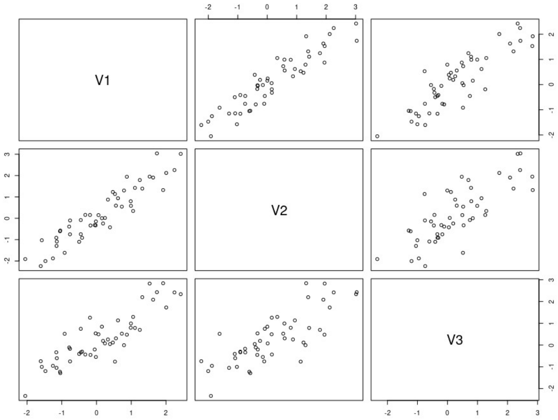
Case 1)
{V1} - {V2} - {V3}



Case 2)
{V1 - V2} - {V3}



Case 3)
{V1 - V2 - V3}



Pearson Correlation matrices

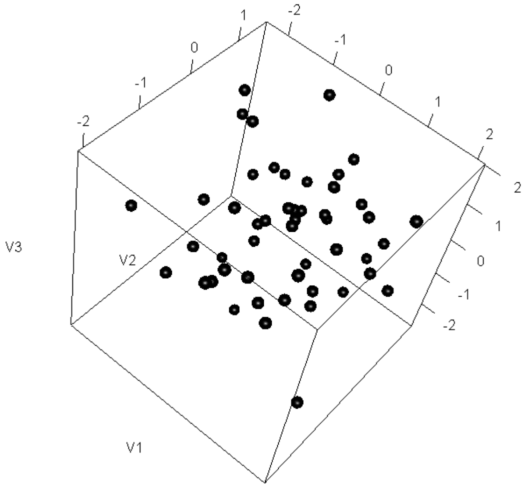
1)	**V1**	**V2**	**V3**
V1	1.00	-0.05	-0.12
V2	-0.05	1.00	0.06
V3	-0.12	0.06	1.00

2)	**V1**	**V2**	**V3**
V1	1.00	0.90	0.08
V2	0.90	1.00	-0.01
V3	0.08	-0.01	1.00

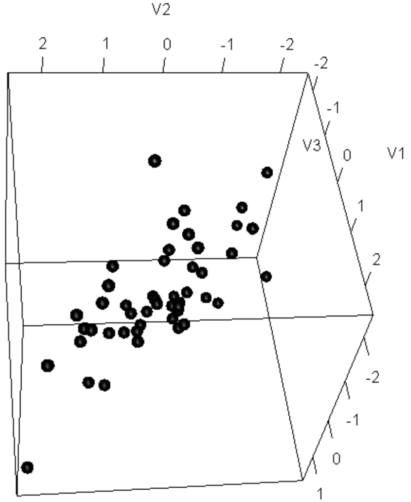
3)	**V1**	**V2**	**V3**
V1	1.00	0.93	0.87
V2	0.93	1.00	0.79
V3	0.87	0.79	1.00

PCA, simulated examples

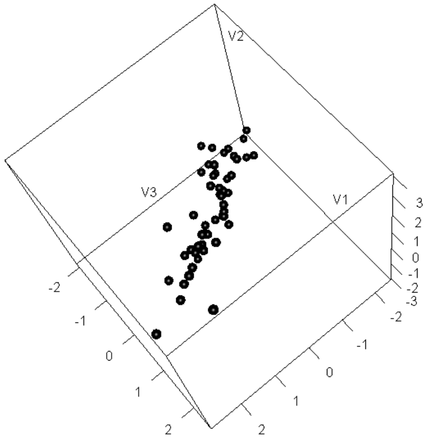
Case 1)



Case 2)



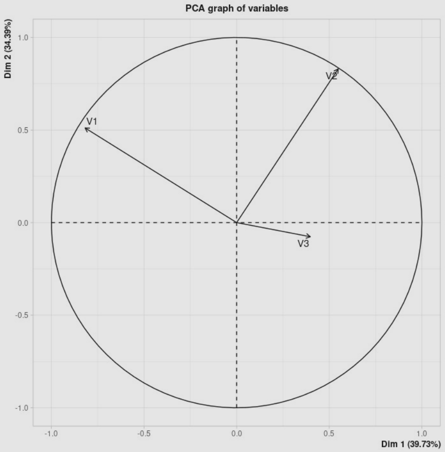
Case 3)

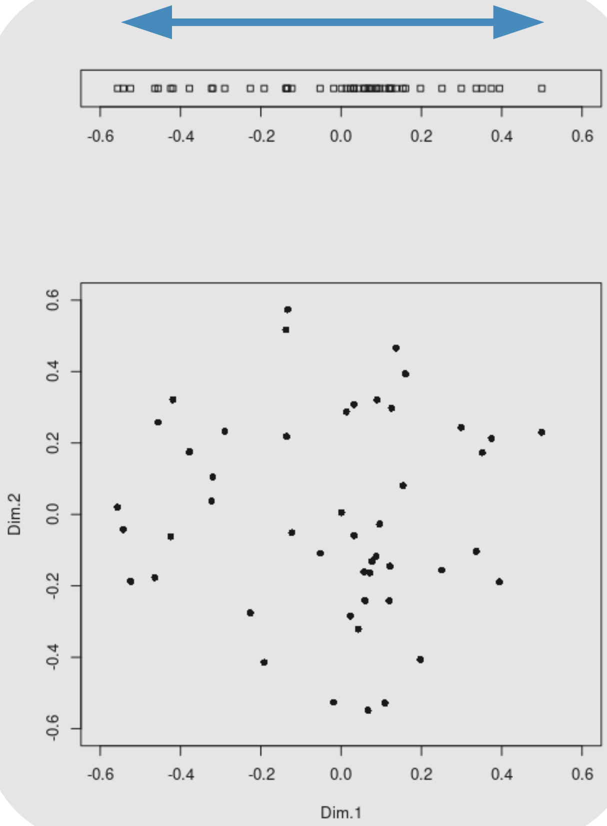


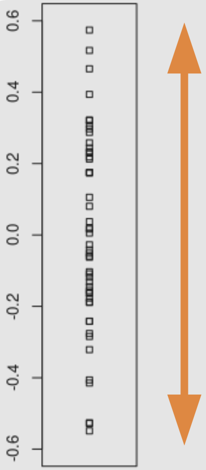
PCA, simulated examples

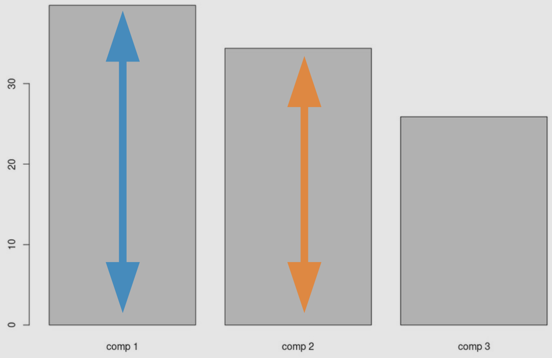
Loadings

	Dim.1	Dim.2	Dim.3
V1	-0.23	0.14	0.07
V2	0.15	0.23	-0.03
V3	0.10	-0.02	0.22



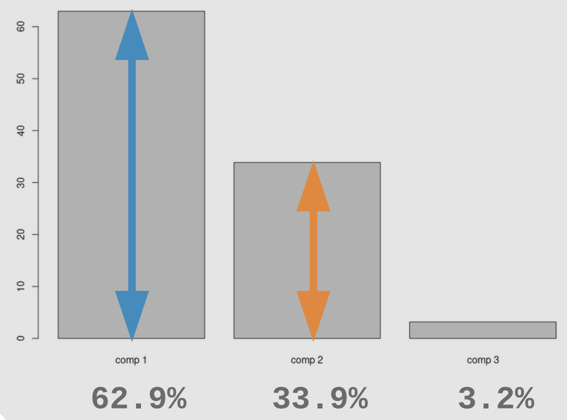






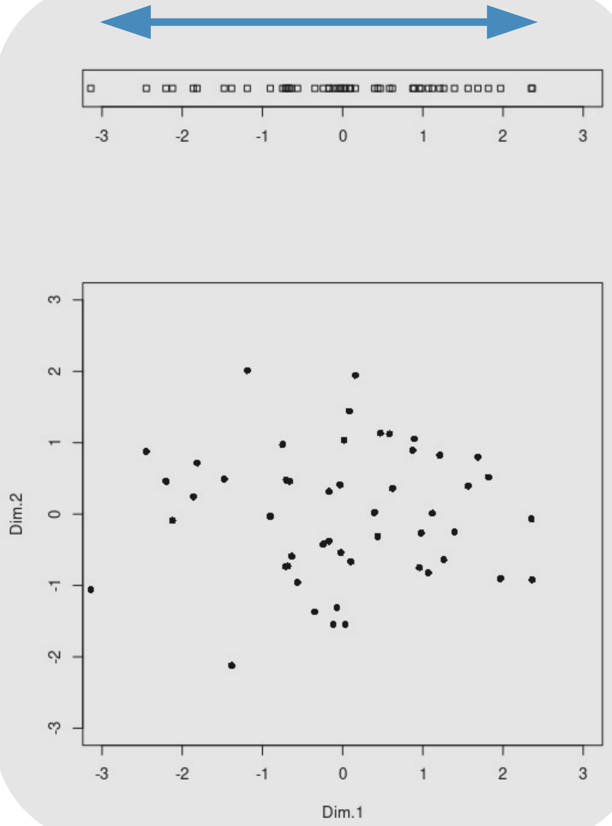
39.7% 34.4% 25.9%

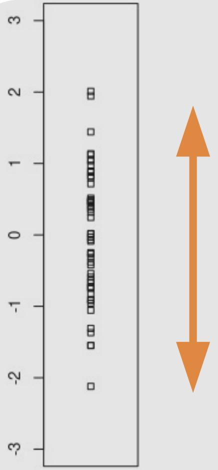
PCA, simulated examples

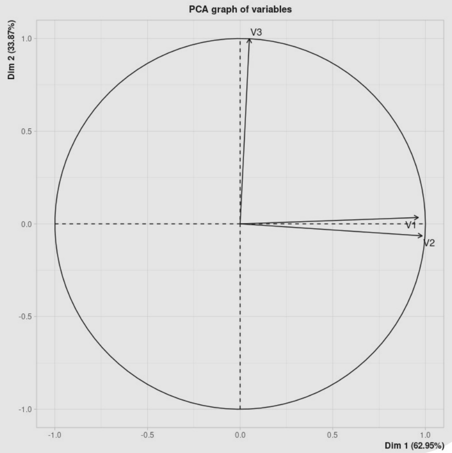


Loadings

	Dim.1	Dim.2	Dim.3
V1	0.77	0.03	0.22
V2	0.97	-0.06	-0.17
V3	0.05	0.91	-0.02



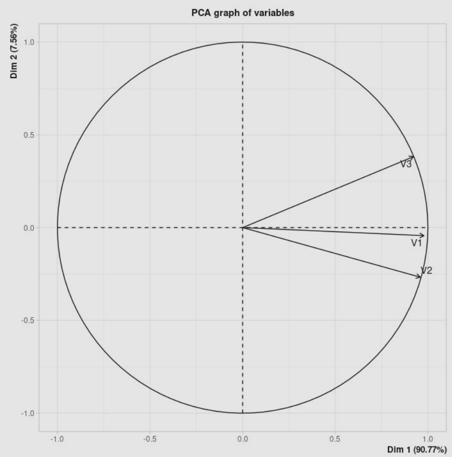


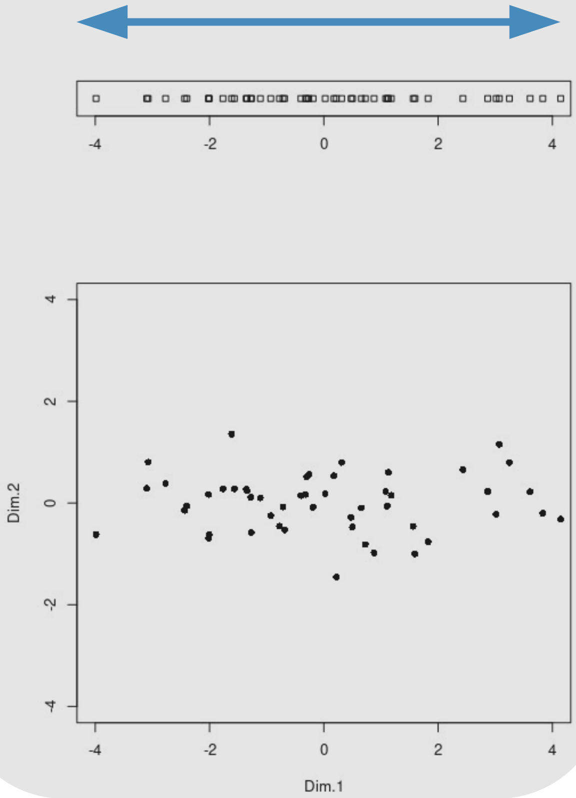


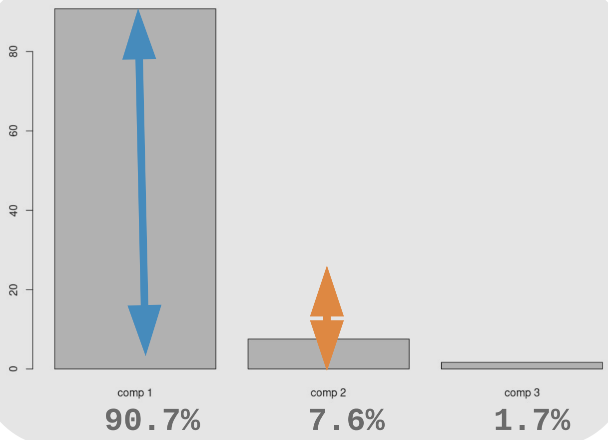
PCA, simulated examples

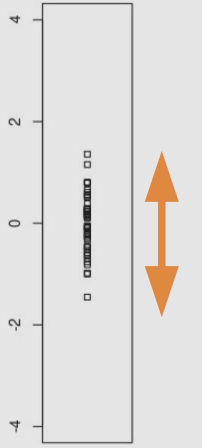
Loadings

	Dim.1	Dim.2	Dim.3
V1	1.07	-0.05	0.22
V2	1.23	-0.34	-0.13
V3	1.07	0.44	-0.07









Extension to integration problems

- **Multivariate unsupervised**

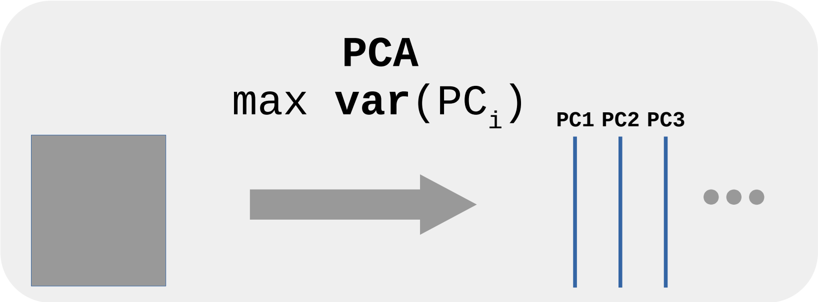
One numerical dataset

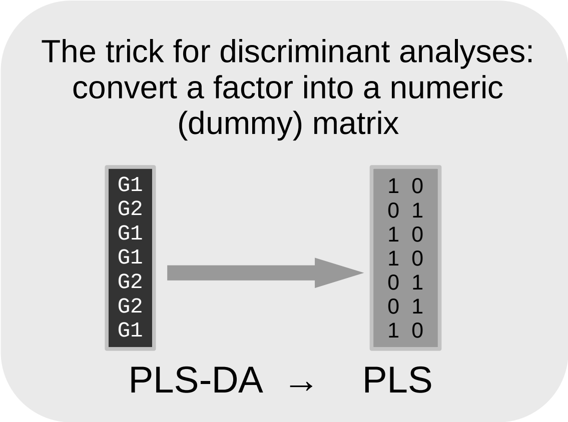
`Principal Component Analysis`

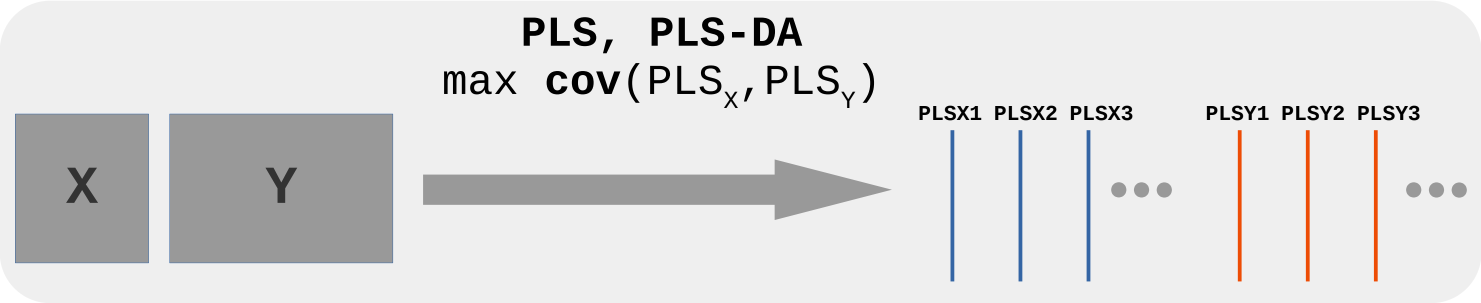
Numerical

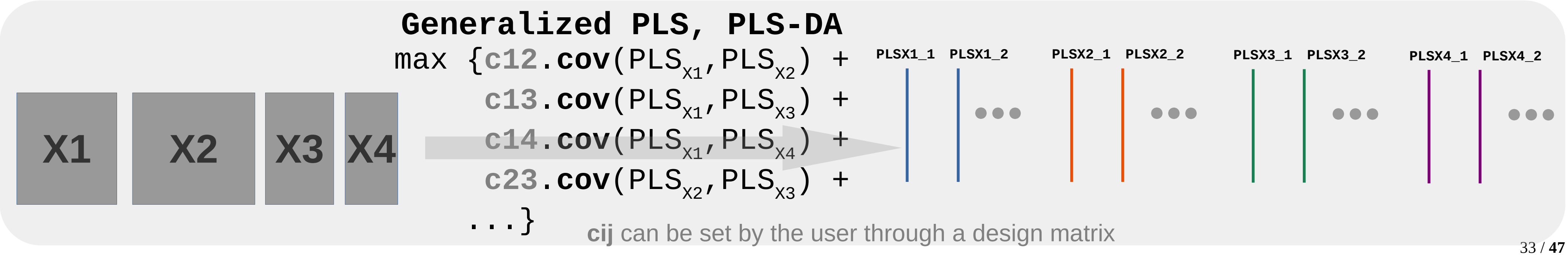
Categorical

Extension to integration problems







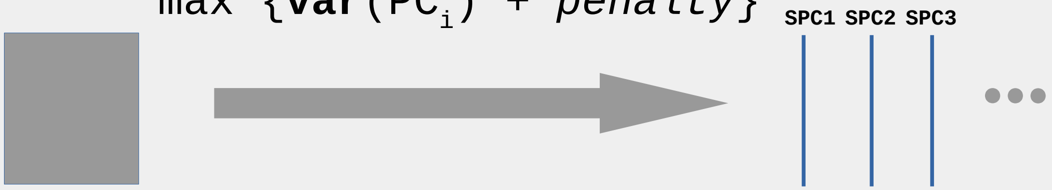


Sparsity

- High throughput experiments: too many variables, noisy or irrelevant depending on the goal aimed
- Some of the variable loadings, among the smallests, are set to 0 thanks to a LASSO (L_1) penalty
- Associated variables are not taken into account when calculating the PCs

Sparse PCA

max {**var**(PC_i) + *penalty*}




```
SPC1 =   0.X1 + a12.X2 + a13.X3 + … +   0.Xp
SCP2 = a21.X1 +   0.X2 +   0.X3 + … + a2p.Xp
...
```

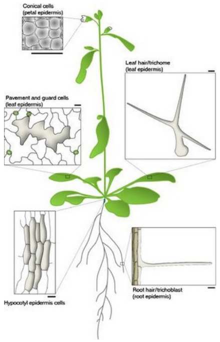
How is global warming affecting plant growth?

WallOmics project

1/ Collect plants on the ground

2/ Gather seeds and grow them in controlled conditions (temperature, light, humidity…) at 2 different temperatures (15°C and 22°C)

3/ Collect different parts of the plant (stem, leaf, rosette...)



4/ Analyze biological samples using high-throughput bio-technologies (DNA sequencing, mass spectrometry...)

5/ Generate very large datasets (thousands of features for each biological sample)

→ need for statistical skills to analyze them

	Pectin_RGI	Pectin_HG	XG	Pectin_linearity	Contribution_RG	RGI_branching
Col.22.1	75.96	60.29	92.88	0.94	0.29	2.70
Col.22.2	63.71	76.68	89.76	1.32	0.17	3.49
Col.22.3	69.05	78.73	103.20	1.28	0.20	2.92
Col.15.1	57.56	43.65	81.75	0.85	0.20	4.95
Col.15.2	79.39	74.34	116.76	1.03	0.16	4.92
Col.15.3	84.36	73.31	123.27	0.96	0.17	5.18
Roch.22.1	89.13	109.42	117.23	1.37	0.20	2.69
Roch.22.2	120.02	138.92	135.48	1.33	0.24	2.16
Roch.22.3	97.46	114.35	130.65	1.33	0.22	2.48
Roch.15.1	91.94	88.57	136.65	1.07	0.19	4.04
Roch.15.2	100.44	96.91	193.22	1.04	0.14	5.95
Roch.15.3	96.42	97.84	179.30	1.09	0.13	6.07
Grip.22.1	97.44	119.20	113.23	1.38	0.21	2.50
Grip.22.2	90.28	88.47	111.65	1.12	0.24	2.76
Grip.22.3	45.95	54.63	58.89	1.29	0.14	4.47
Grip.15.1	77.22	72.26	99.00	1.01	0.14	6.26
Grip.15.2	80.55	77.47	122.85	1.04	0.14	6.08
Grip.15.3	86.40	82.43	132.43	1.03	0.13	6.24

WallOmics project: datasets

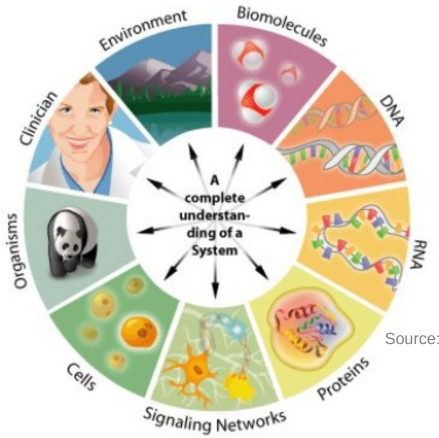
- R package WallomicsData

`CRAN.R-project.org/package=WallomicsData`

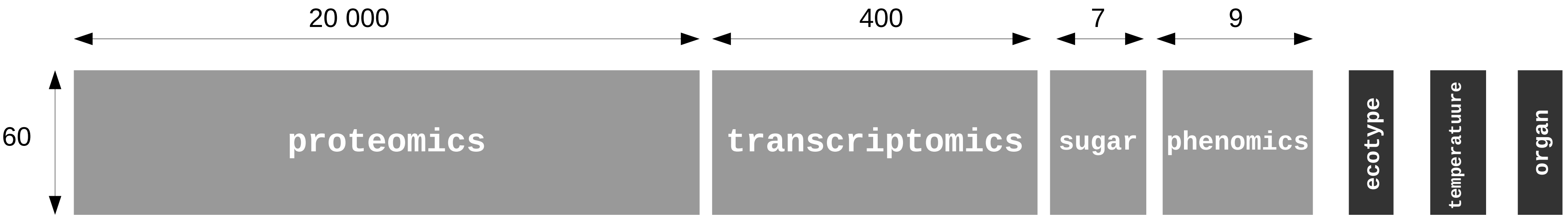
- **60** samples A. thaliana:
  - **5** ecotypes (Col, Grip, Hern, Roc, Hosp)
  - **2** temperatures (low, high)
  - **2** organs (stem, rosette)
  - **3** replicates
- **4** tables: proteomics, transcriptomics, metabolomics (sugar), phenomics

Generally, data integration can be defined as the process of combining data residing in diverse sources to provide users with a comprehensive view of such data. There is no universal approach to data integration, and many techniques are still evolving.

Schneider, M. V., & Jimenez, R. C. (2012). Teaching the Fundamentals of Biological Data Integration Using Classroom Games. PLoS Computational Biology, 8(12)



Source: `la-biologia6.webnode.es`



WallOmics project: one specific question

Can We Determine a Multi-omics Signature to Classify Ecotypes on the Basis of Floral Stem Data?

- **Multi-omics**: consider all the datasets (proteomics, transcriptomics, metabolomics, phenomics → *multi-block analysis*
- **Signature**: select the most relevant variables inside each dataset → *sparsity*
- **Classify ecotypes**: supervised method → *Discriminant Analysis*
- **Floral stem**: filter data ‘organ = stem’

WallOmics project: method

Multi-Block
Sparse
Projection to Latent Structures
Discriminant Analysis

Bioinformatics, 35(17), 2019, 3055–3062
doi: 10.1093/bioinformatics/bty1054
Advance Access Publication Date: 18 January 2019
Original Paper

Systems biology

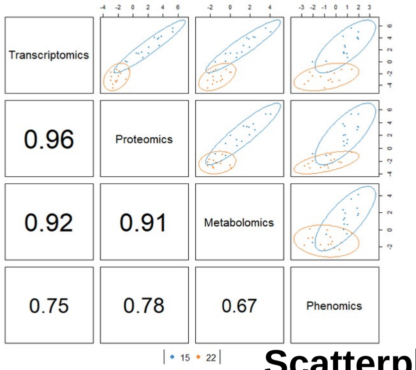
DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays

Amrit Singh[1], Casey P. Shannon[1], Benoît Gautier[2], Florian Rohart[3], Michaël Vacher[4], Scott J. Tebbutt[1] and Kim-Anh Lê Cao[5,*]

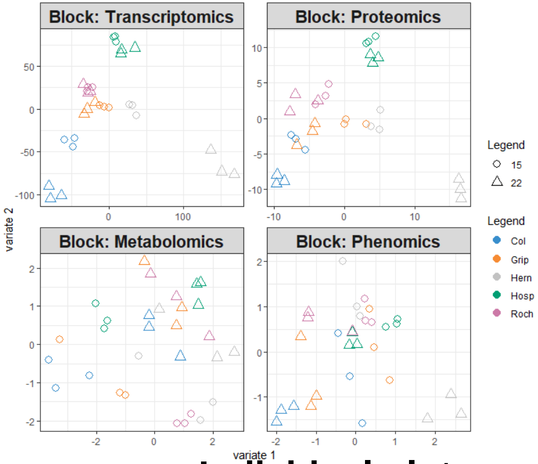
$$\max_{a_h^{(1)},\dots,a_h^{(Q)}} \sum_{i,j=1,\, i\neq j}^{Q} c_{i,j}\ \mathrm{cov}\left(X_h^{(i)} a_h^{(i)}, X_h^{(j)} a_h^{(j)}\right),$$

$$\text{s.t. } ||a_h^{(q)}||_2 = 1 \ \text{ and } \ ||a_h^{(q)}||_1 \leq \lambda^{(q)} \ \textit{for all } 1 \leq q \leq Q$$

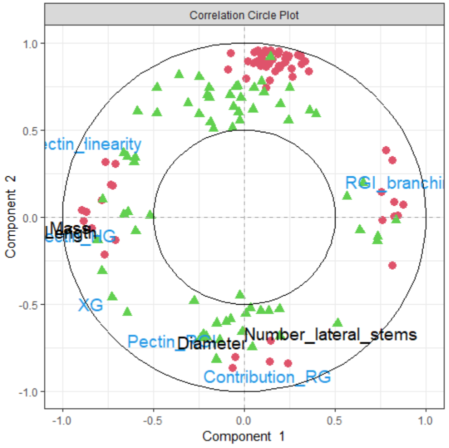
WallOmics project: results



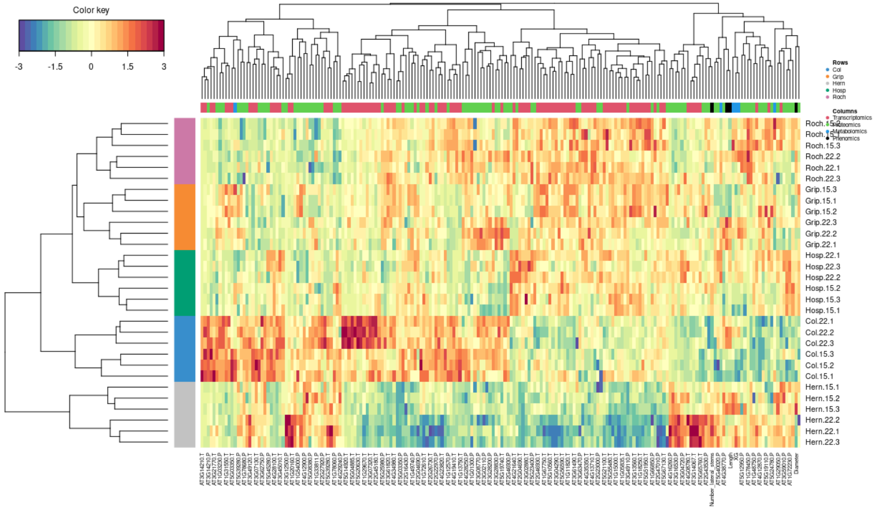
Scatterplot matrix



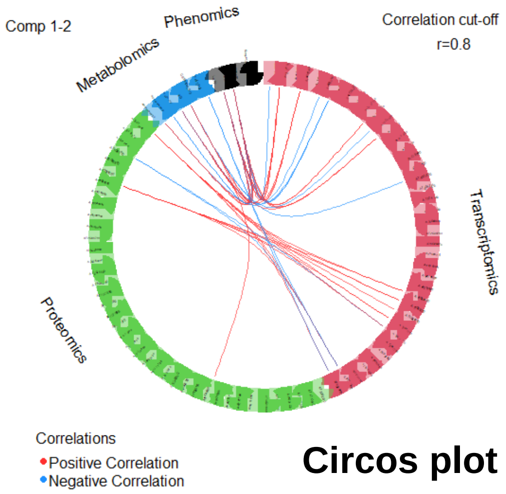
Individual plot



Correlation circle plot



Clustered Image Map / heatmap



Circos plot

WallOmics project: publication

Volume 22, Issue 3

JOURNAL ARTICLE

A powerful framework for an integrative study with heterogeneous omics data: from univariate statistics to multi-block analysis Get access ›

Harold Duruflé, Merwann Selmani, Philippe Ranocha, Elisabeth Jamet, Christophe Dunand ✉, Sébastien Déjean ✉

Plant Science
Volume 263, October 2017, Pages 183-193
ELSEVIER

Cell wall modifications of two *Arabidopsis thaliana* ecotypes, Col and Sha, in response to sub-optimal growth conditions: An integrative study

Harold Duruflé [a, 1], Vincent Hervé [a, 1], Philippe Ranocha [a], Thierry Balliau [b, c], Michel Zivy [b, c], Josiane Chourré [a], Hélène San Clemente [a], Vincent Burlat [a], Cécile Albenne [a], Sébastien Déjean [d], Elisabeth Jamet [a], Christophe Dunand [a]

- **Harold**: PhD student, vegetal biology
- **Merwann**: intern, applied mathematics
- **Philippe**: researcher biology
- **Elisabeth**: professor in biology
- **Christophe**: professor in biology, Harold’s supervisor
- **Sébastien** : statistician, Harold’s co-supervisor

Phenotypic Trait Variation as a Response to Altitude-Related Constraints in Arabidopsis Populations

frontiers
Frontiers in Plant Science

Harold Duruflé[1†], Philippe Ranocha[1†], Duchesse Lacour Mbadinga Mbadinga[1], Sébastien Déjean[2], Maxime Bonhomme[1], Hélène San Clemente [1], Sébastien Viudes[1], Ali Eljebbawi[1], Valerie Delorme-Hinoux[3,4], Julio Sáez-Vásquez[3,4], Jean-Philippe Reichheld[3,4], Nathalie Escaravage[5], Monique Burrus[5] and Christophe Dunand[1*]

cells

Submit to this Journal

Review for this Journal

Edit a Special Issue

Article Menu

Open Access | Feature Paper | Article

An Integrative Study Showing the Adaptation to Sub-Optimal Growth Conditions of Natural Populations of *Arabidopsis thaliana*: A Focus on Cell Wall Changes

by Harold Duruflé [1], Philippe Ranocha [1], Thierry Balliau [2], Michel Zivy [2], Cécile Albenne [1], Vincent Burlat [1], Sébastien Déjean [3,*], Elisabeth Jamet [1] and Christophe Dunand [1,*]

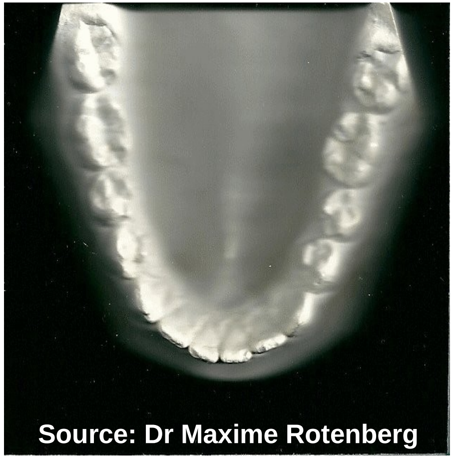
Other problems in the real world where maths can help

Maths help... to give you a shining smile

Does wearing dental braces for 18 months really work?

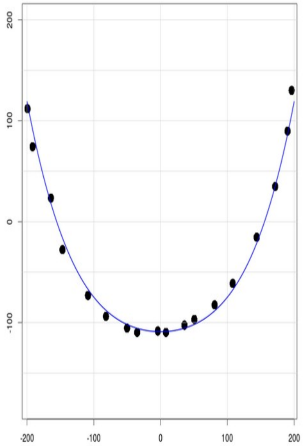
For a shining smile...

… you must go
through this...



Source: Dr Maxime Rotenberg

… and maybe also that!



Modeling the dental arch with a 4 degree polynom without odd degres terms (for axial symetry)

$$Y = b_0 + b_2X^2 + b_4X^4$$

M. Rotenberg, Modélisation de la forme d'arcade dentaire de jeunes adultes `www.theses.fr/1996TOU30012`

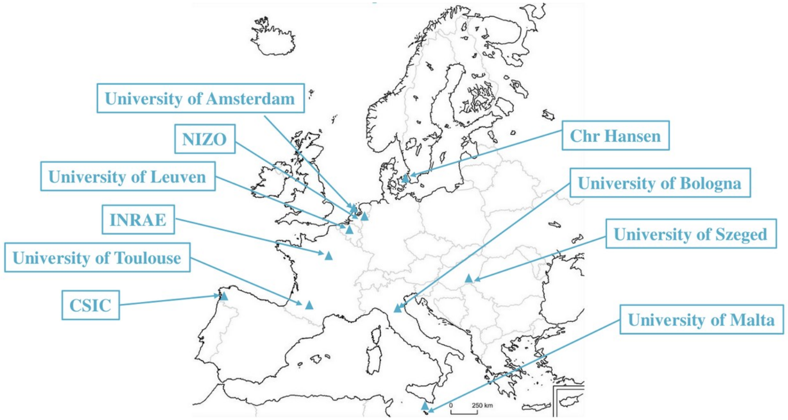
Maths help... to understand cheese ripening

What are the microbiological mechanisms involved in the cheese ripening process?

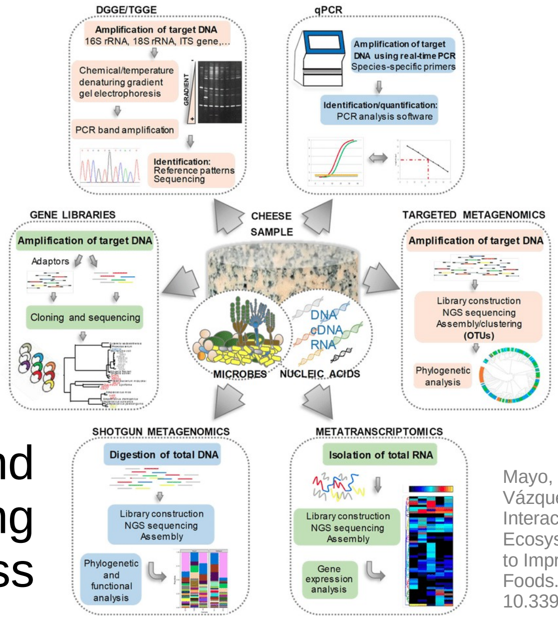
Complex microbial Ecosystems MUltiScale modElling: mechanistic and data driven approaches integration

www.itn-emuse.com

10 Europeans partners...



... to better understand the cheese ripening process

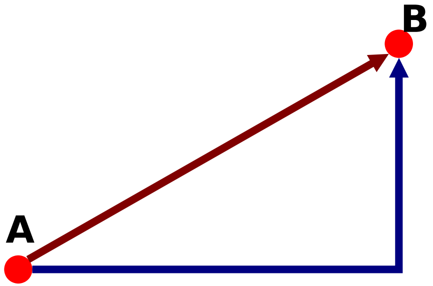


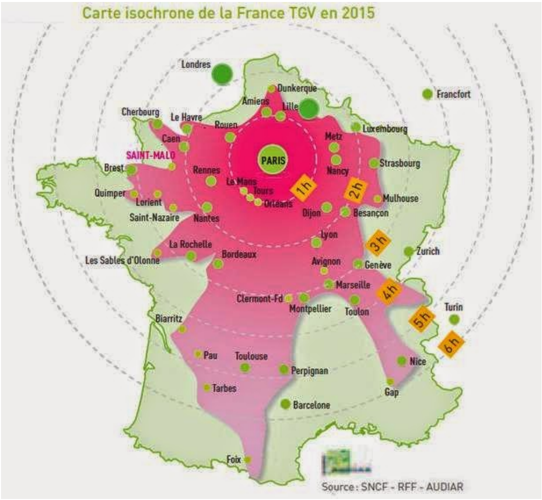
Mayo, Rodríguez Álvarez, Vázquez, Flórez (2021). Microbial Interactions within the Cheese Ecosystem and Their Application to Improve Quality and Safety. Foods. 10. 602. 10.3390/foods10030602.

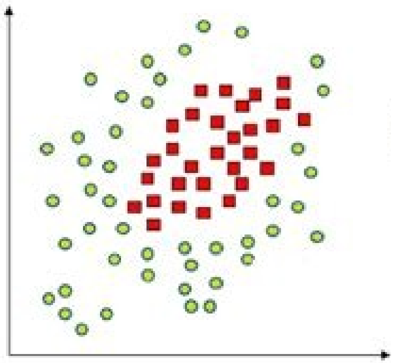
“This presentation is part of a project that has received funding from the European Union’s Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 956126”.

Maths help... to understand cheese ripening

PhD thesis (ongoing): Kernel approaches for the integration of biological data from heterogeneous sources

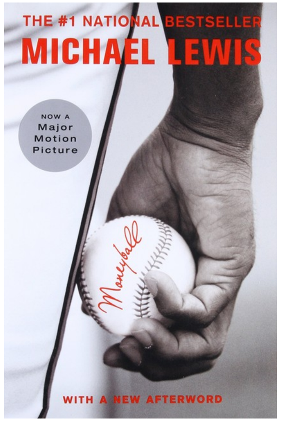


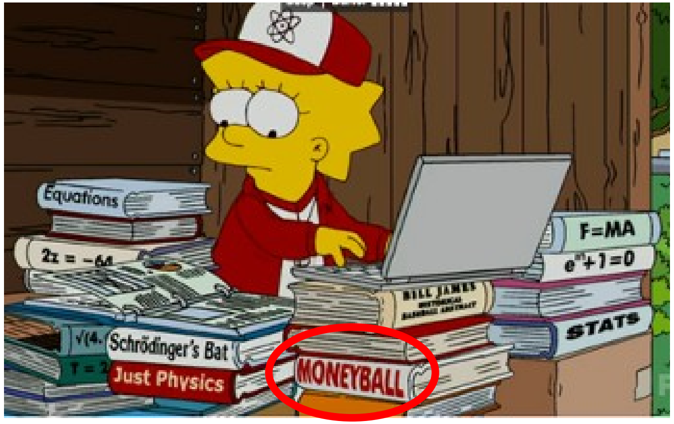


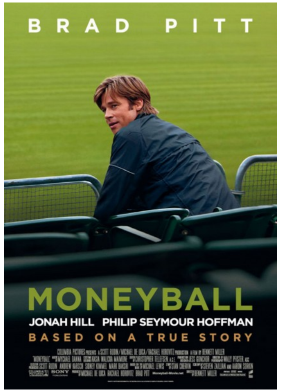


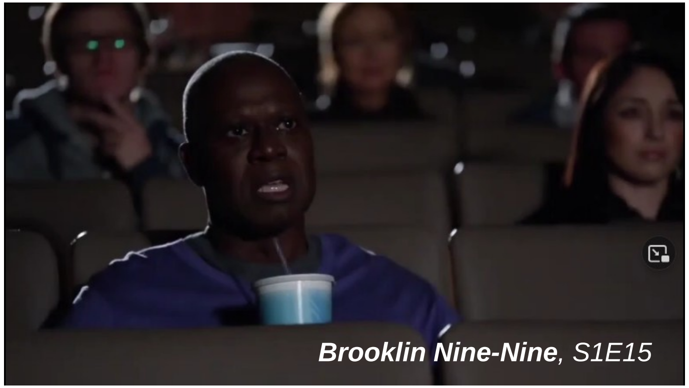
Source : `ichi.pro/de/was-ist-der-kernel-trick-warum-ist-es-wichtig-167853994055197`

Maths help... to optimize performance in sport









***Brooklin Nine-Nine**, S1E15*

- Recruit new players
- Prevent injury
- Model collective behaviour
- Identify optimal strategies
- ...

Quand les maths se mêlent de sport

15.03.2016, par Laure Cailloce
Mis à jour le 07.07.2016

Un modèle développé au Laboratoire de mathématiques de Versailles permet aux athlètes de courir la course optimale, en fonction de la distance à parcourir et de quelques paramètres clés : capacité pulmonaire, apports énergétiques, ou encore, force maximale de propulsion...

lejournal.cnrs.fr/articles/quand-les-maths-se-melent-de-sport

Other domains

ELSEVIER
Journal of Cereal Science
Volume 107, September 2022, 103533

Differences in bread protein digestibility traced to wheat cultivar traits

Mélanie Lavoignat [a, b], Sylvain Denis [c], Annie Faye [a], Laura Halupka [d], Sibille Perrochon [a], Larbi Rhazi [e], Pascal Giraudeau [g], Sébastien Déjean [f], Gérard Branlard [a], Emmanuelle Bancel [a, 1], Catherine Ravel [a] [1]

JOURNAL OF VOICE — THE VOICE FOUNDATION
FULL LENGTH ARTICLE | VOLUME 31, ISSUE 2, P261.E33-261.E38, MARCH 01, 2017

Vocal Problems in Sports and Fitness Instructors: A Study of Prevalence, Risk Factors, and Need for Prevention in France

Lionel Fontan • Marie Fraval • Anne Michon • Sébastien Déjean • Muriel Welby-Gieusse

Published: August 12, 2016 • DOI: https://doi.org/10.1016/j.jvoice.2016.04.014

The American Journal of CLINICAL NUTRITION
A journal of the American Society for Nutrition

Gene expression profiling of human skeletal muscle in response to stabilized weight loss [1]

Dominique Larrouy, Pierre Barbe, Carine Valle, Sébastien Déjean, Véronique Pelloux, Claire Thalamas, Jean-Philippe Bastard, Anne Le Bouil, Bertrand Diquet, Karine Clément, Dominique Langin, Nathalie Viguerie

cancers

Quantitative Analysis of Cell Aggregation Dynamics Identifies HDAC Inhibitors as Potential Regulators of Cancer Cell Clustering

by Fabien Gava [1], Julie Pignolet [1], Sébastien Déjean [2], Odile Mondésert [1], Renaud Morin [3], Joseph Agossa [1,2], Bernard Ducommun [1,4] and Valérie Lobjois [1,5,*]

ELSEVIER
Water Research
Volume 140, 1 September 2018, Pages 24-33

Support media can steer methanogenesis in the presence of phenol through biotic and abiotic effects

Simon Poirier [a], Sébastien Déjean [b], Olivier Chapleur [a]

Show more

Add to Mendeley Share Cite

frontiers in Psychiatry
ORIGINAL RESEARCH
published: 31 March 2020
doi: 10.3389/fpsyt.2020.00165

Urinary Amine and Organic Acid Metabolites Evaluated as Markers for Childhood Aggression: The ACTION Biomarker Study

OPEN ACCESS
Edited by:
Marieke Klein,
Radboud University Nijmegen Medical Centre, Netherlands
Reviewed by:

Fiona A. Hagenbeek [1,2*†], Peter J. Roetman [3†], René Pool [1,2], Cornelis Kluft [4], Amy C. Harms [5,6], Jenny van Dongen [1,2], Olivier F. Colins [3,7], Simone Talens [4], Catharina E. M. van Beijsterveldt [1], Marjolein M. L. J. Z. Vandenbosch [1], Eveline L. de Zeeuw [1,2], Sébastien Déjean [8], Vassilios Fanos [9], Erik A. Ehli [10], Gareth E. Davies [10], Jouke Jan Hottenga [1], Thomas Hankemeier [5,6], Meike Bartels [1,2,11],

Knowl Inf Syst (2012) 30:693–713
DOI 10.1007/s10115-011-0391-7
REGULAR PAPER

How many performance measures to evaluate information retrieval systems?

Alain Baccini · Sébastien Déjean · Laetitia Lafage · Josiane Mothe

Received: 15 September 2010 / Revised: 4 January 2011 / Accepted: 30 January 2011 / Published online: 12 April 2011

When Bigger Is Better: 3D RNA Profiling of the Developing Head in the Catshark *Scyliorhinus canicula*

frontiers — Frontiers in Cell and Developmental Biology

Hélène Mayeur[1], Maxence Lanoizelet[1], Aurélie Quillien, Arnaud Menuet, Léo Michel, Kyle John Martin[4], Sébastien Dejean[5], Patrick Blader[2], Sylvie Mazan[1*†] and Ronan Lagadec[1†]