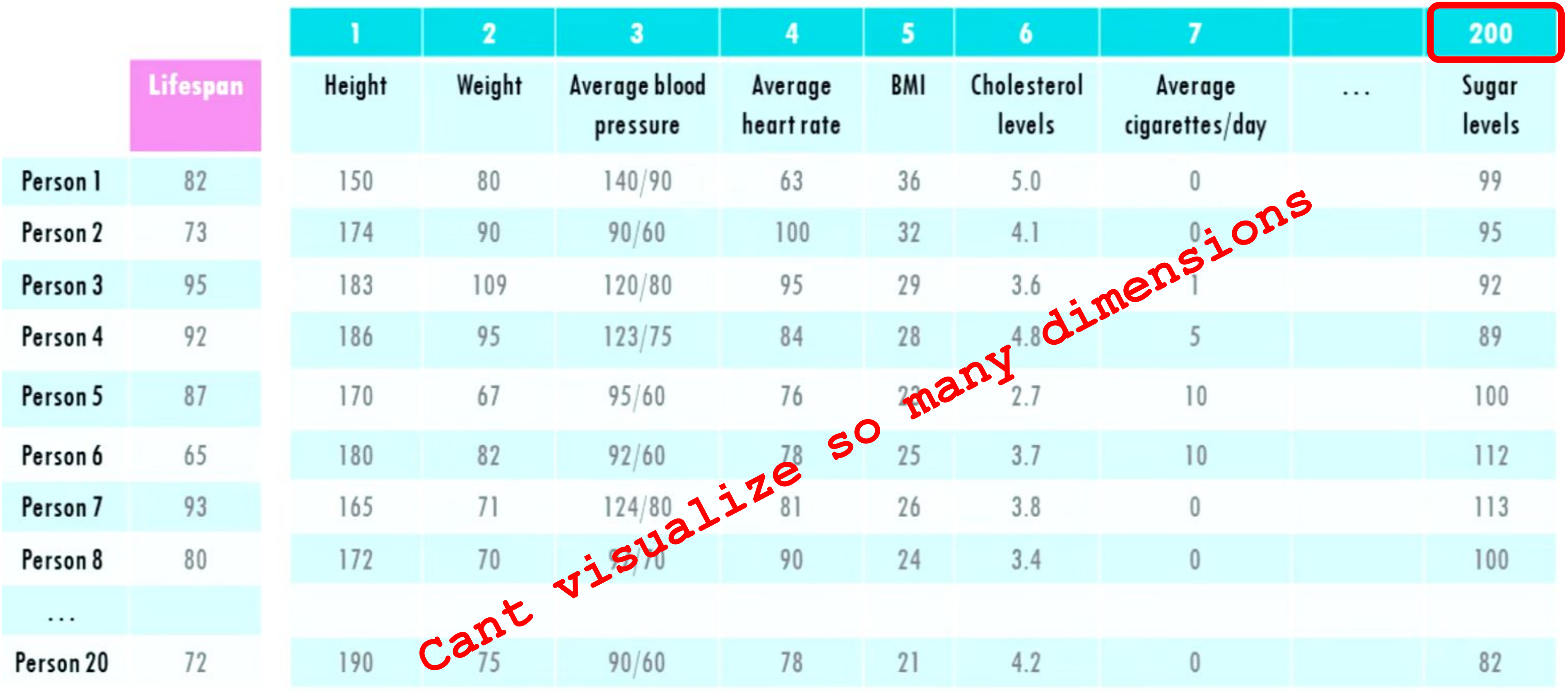
Principal Component Analysis

— What & Why —

Source:

https://youtu.be/5vgP05YpKdE?si=Uiw3hlALHmoSVVzK | Biostatsquid

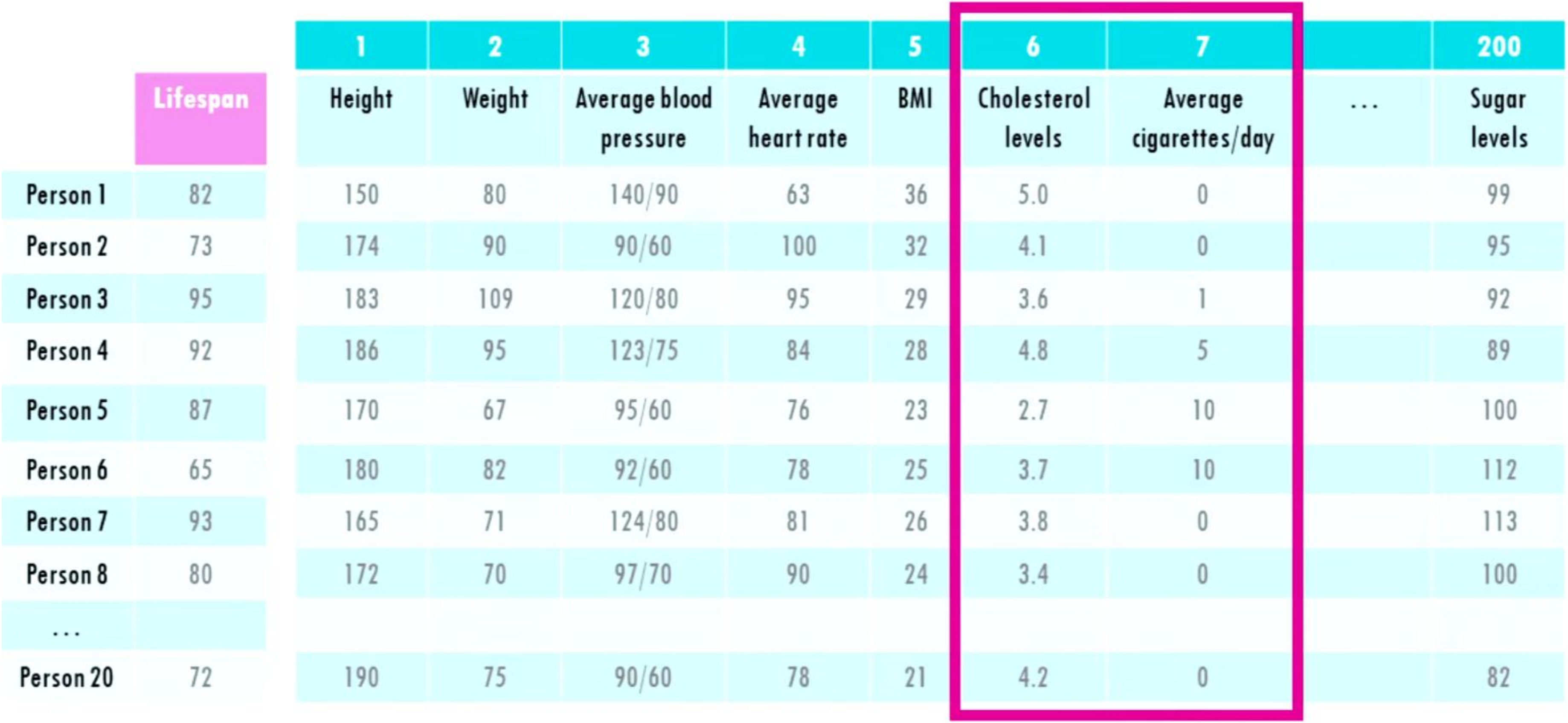
Goal: What factors contribute to a longer/shorter lifespan?



		1	2	3	4	5	6	7		200
	Lifespan	Height	Weight	Average blood pressure	Average heart rate	BMI	Cholesterol levels	Average cigarettes/day	. . .	Sugar levels
Person 1	82	150	80	140/90	63	36	5.0	0		99
Person 2	73	174	90	90/60	100	32	4.1	0		95
Person 3	95	183	109	120/80	95	29	3.6	1		92
Person 4	92	186	95	123/75	84	28	4.8	5		89
Person 5	87	170	67	95/60	76	23	2.7	10		100
Person 6	65	180	82	92/60	78	25	3.7	10		112
Person 7	93	165	71	124/80	81	26	3.8	0		113
Person 8	80	172	70	9[illegible]/70	90	24	3.4	0		100
. . .										
Person 20	72	190	75	90/60	78	21	4.2	0		82

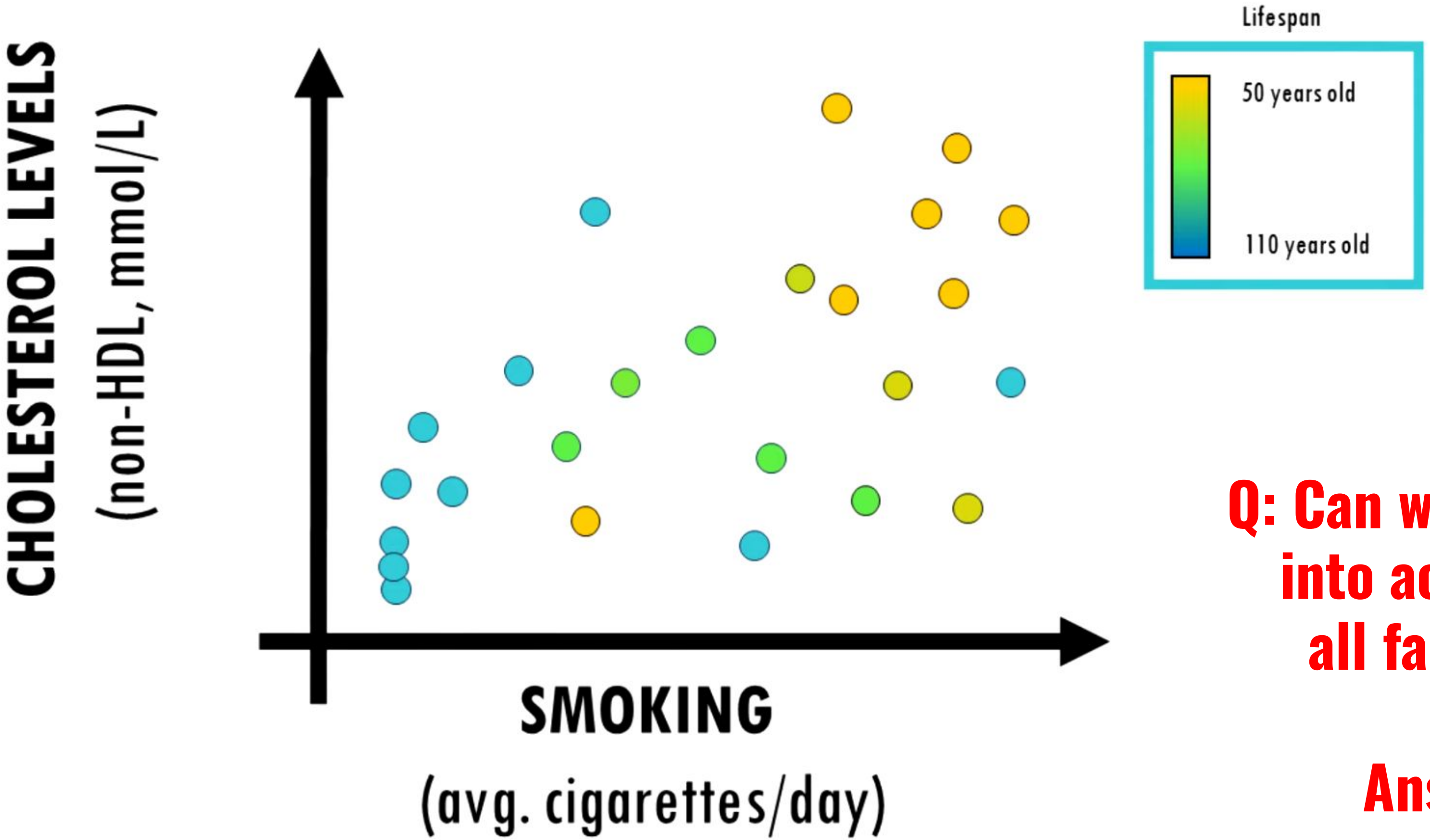
Cant visualize so many dimensions

Goal: What factors contribute to a longer/shorter lifespan?



		1	2	3	4	5	6	7		200
	Lifespan	Height	Weight	Average blood pressure	Average heart rate	BMI	Cholesterol levels	Average cigarettes/day	...	Sugar levels
Person 1	82	150	80	140/90	63	36	5.0	0		99
Person 2	73	174	90	90/60	100	32	4.1	0		95
Person 3	95	183	109	120/80	95	29	3.6	1		92
Person 4	92	186	95	123/75	84	28	4.8	5		89
Person 5	87	170	67	95/60	76	23	2.7	10		100
Person 6	65	180	82	92/60	78	25	3.7	10		112
Person 7	93	165	71	124/80	81	26	3.8	0		113
Person 8	80	172	70	97/70	90	24	3.4	0		100
...										
Person 20	72	190	75	90/60	78	21	4.2	0		82

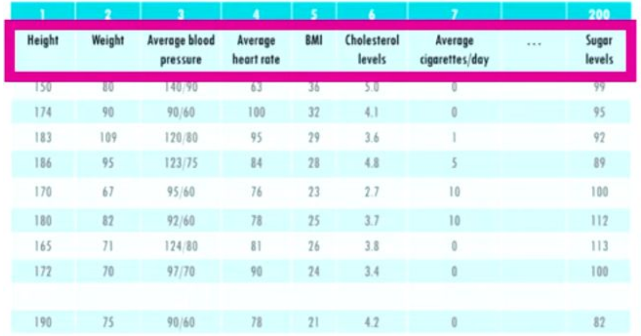
Goal: What factors contribute to a longer/shorter lifespan?



Q: Can we take into account all factors?

Ans: PCA

Goal: What factors contribute to a longer/shorter lifespan?



1	2	3	4	5	6	7		200
Height	Weight	Average blood pressure	Average heart rate	BMI	Cholesterol levels	Average cigarettes/day	...	Sugar levels
150	80	140/90	63	36	5.0	0		99
174	90	90/60	100	32	4.1	0		95
183	109	120/80	95	29	3.6	1		92
186	95	123/75	84	28	4.8	5		89
170	67	95/60	76	23	2.7	10		100
180	82	92/60	78	25	3.7	10		112
165	71	124/80	81	26	3.8	0		113
172	70	97/70	90	24	3.4	0		100
190	75	90/60	78	21	4.2	0		82

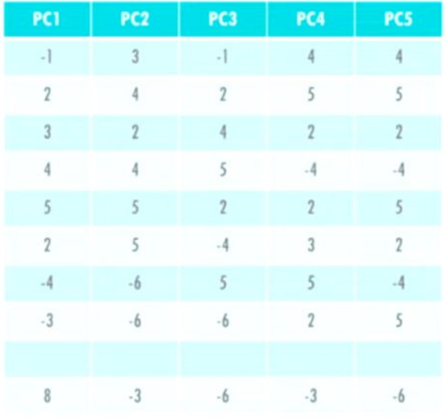
Principal
Component
Analysis

THE FIRST FEW PRINCIPAL

COMPONENTS HOLD

MOST OF THE INFORMATION

OF THE DATASET

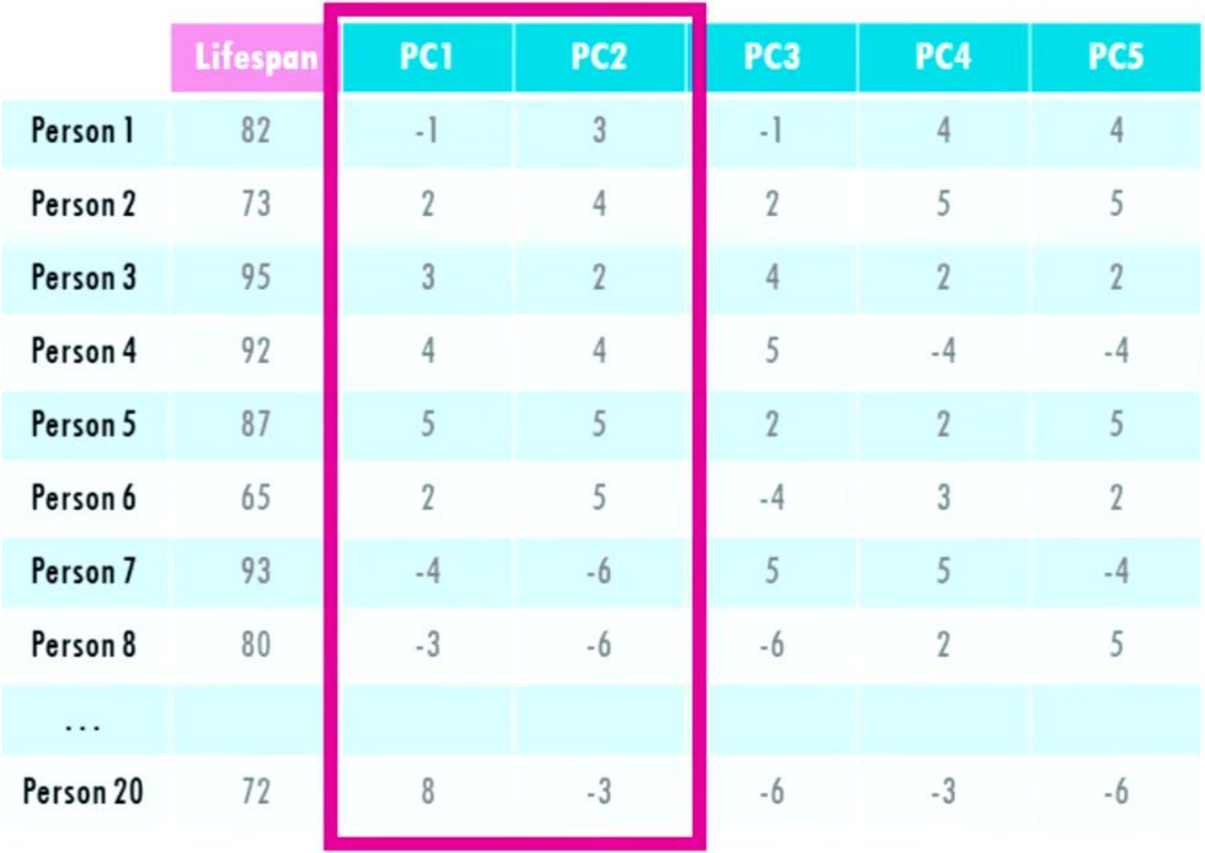


PC1	PC2	PC3	PC4	PC5
-1	3	-1	4	4
2	4	2	5	5
3	2	4	2	2
4	4	5	-4	-4
5	5	2	2	5
2	5	-4	3	2
-4	-6	5	5	-4
-3	-6	-6	2	5
8	-3	-6	-3	-6

5 PRINCIPAL

COMPONENTS

Goal: What factors contribute to a longer/shorter lifespan?

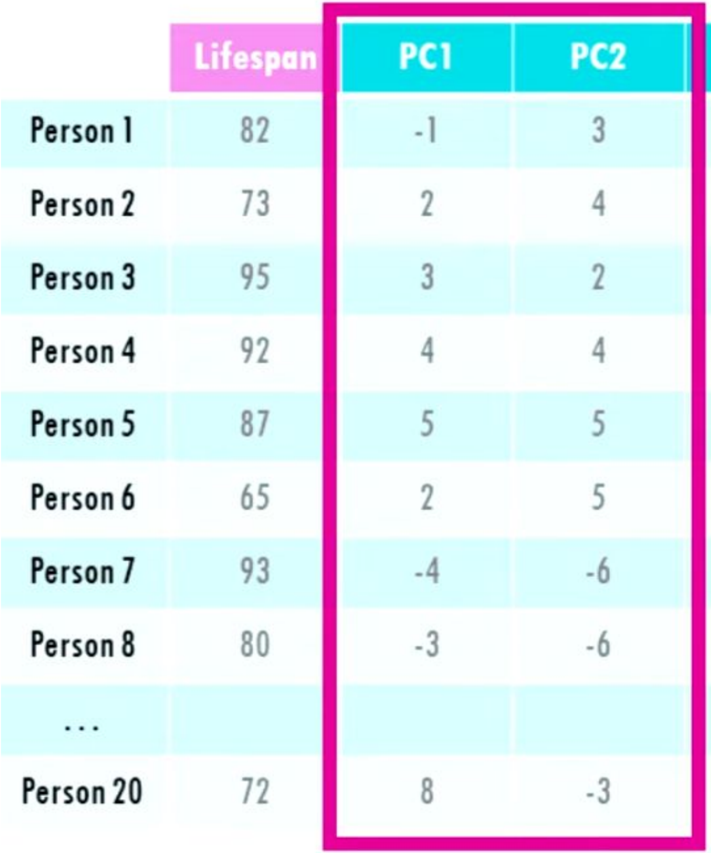


	Lifespan	PC1	PC2	PC3	PC4	PC5
Person 1	82	-1	3	-1	4	4
Person 2	73	2	4	2	5	5
Person 3	95	3	2	4	2	2
Person 4	92	4	4	5	-4	-4
Person 5	87	5	5	2	2	5
Person 6	65	2	5	-4	3	2
Person 7	93	-4	-6	5	5	-4
Person 8	80	-3	-6	-6	2	5
...						
Person 20	72	8	-3	-6	-3	-6

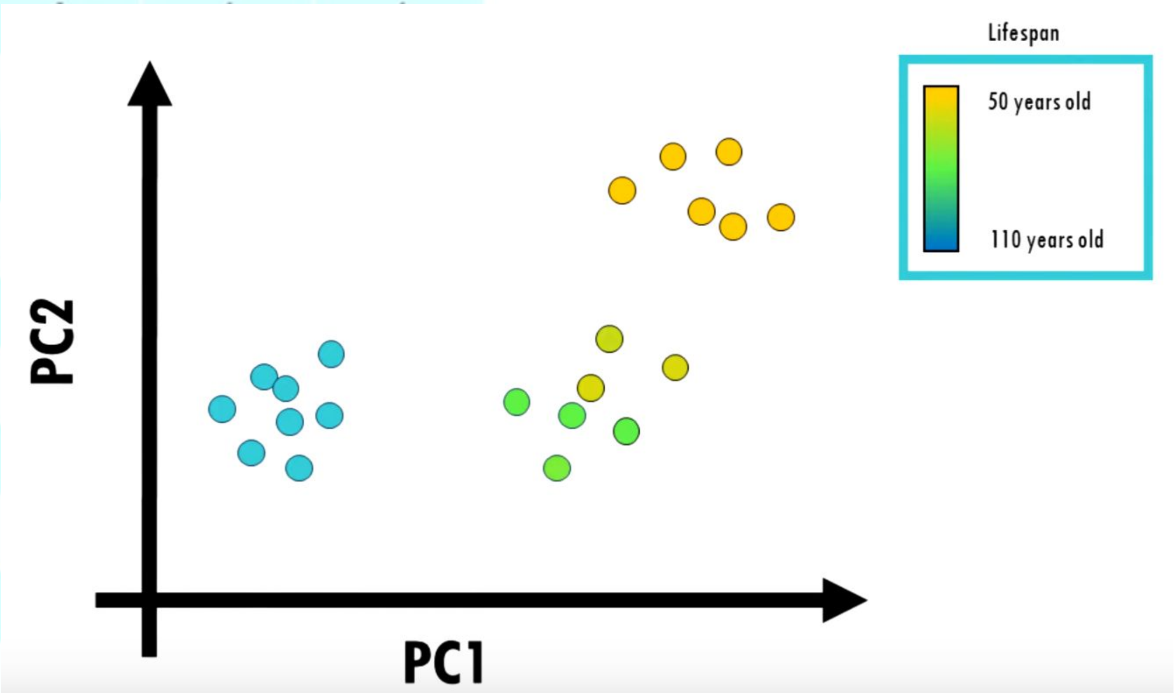
PCA are ranked from most important to least important

PC1 > PC2 > PC3 > PC4 > ...

Goal: What factors contribute to a longer/shorter lifespan?



	Lifespan	PC1	PC2	PC3	PC4	PC5
Person 1	82	-1	3			
Person 2	73	2	4			
Person 3	95	3	2			
Person 4	92	4	4			
Person 5	87	5	5			
Person 6	65	2	5			
Person 7	93	-4	-6			
Person 8	80	-3	-6			
...						
Person 20	72	8	-3			

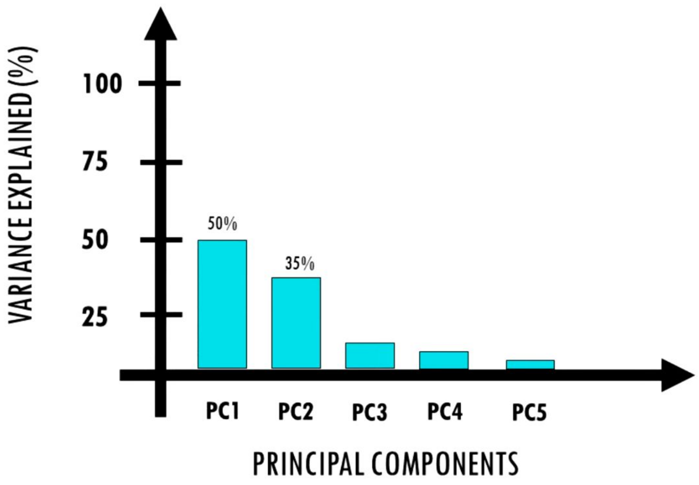


What about the other PCs?

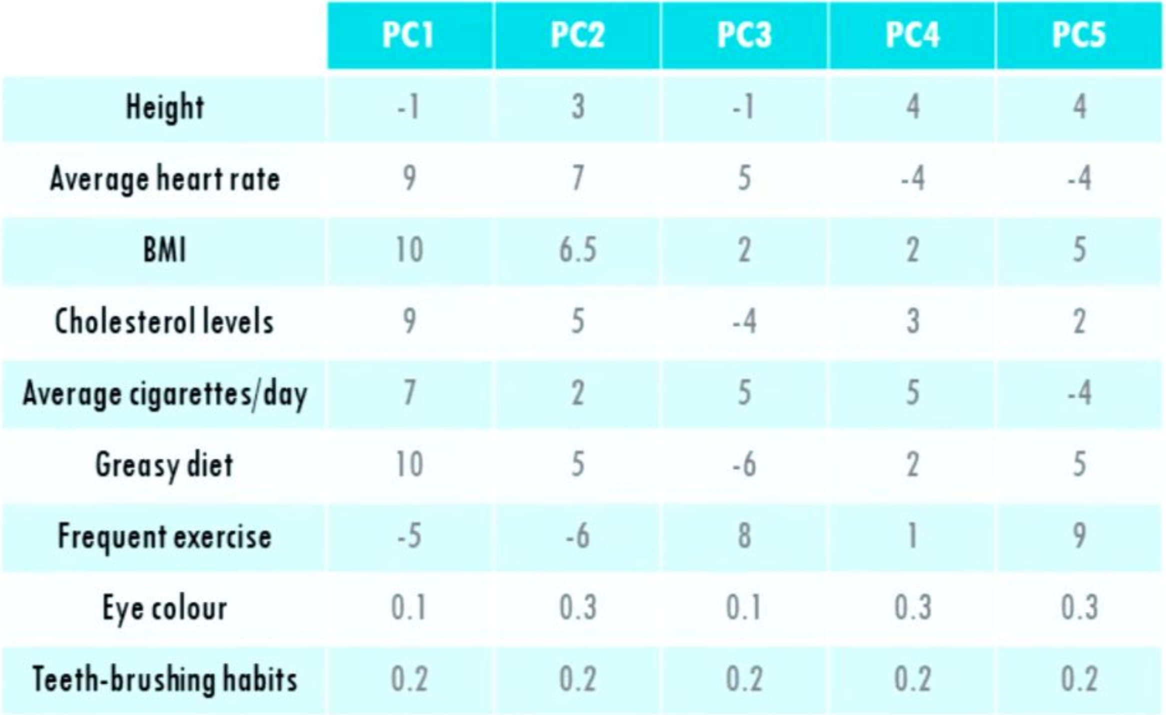
	Lifespan	PC1	PC2	PC3	PC4	PC5
Person 1	82	-1	3	-1	4	4
Person 2	73	2	4	2	5	5
Person 3	95	3	2	4	2	2
Person 4	92	4	4	5	?	-4
Person 5	87	5	5	2		5
Person 6	65	2	5	-4	3	2
Person 7	93	-4	-6	5	5	-4
Person 8	80	-3	-6	-6	2	5
...						
Person 20	72	8	-3	-6	-3	-6

IDEALLY, WE WANT TO GET AROUND 90% VARIANCE WITH JUST 2 TO 3 PCS SO THAT ENOUGH INFORMATION IS RETAINED WHILE WE CAN STILL VISUALIZE OUR DATA ON A PLOT.

SCREE PLOT

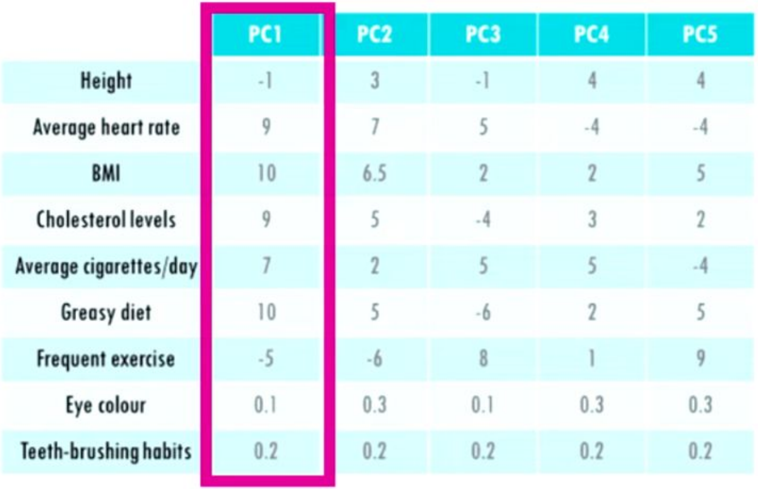


LOADINGS INDICATE THE CONTRIBUTION OF THE VARIABLES TO EACH PC

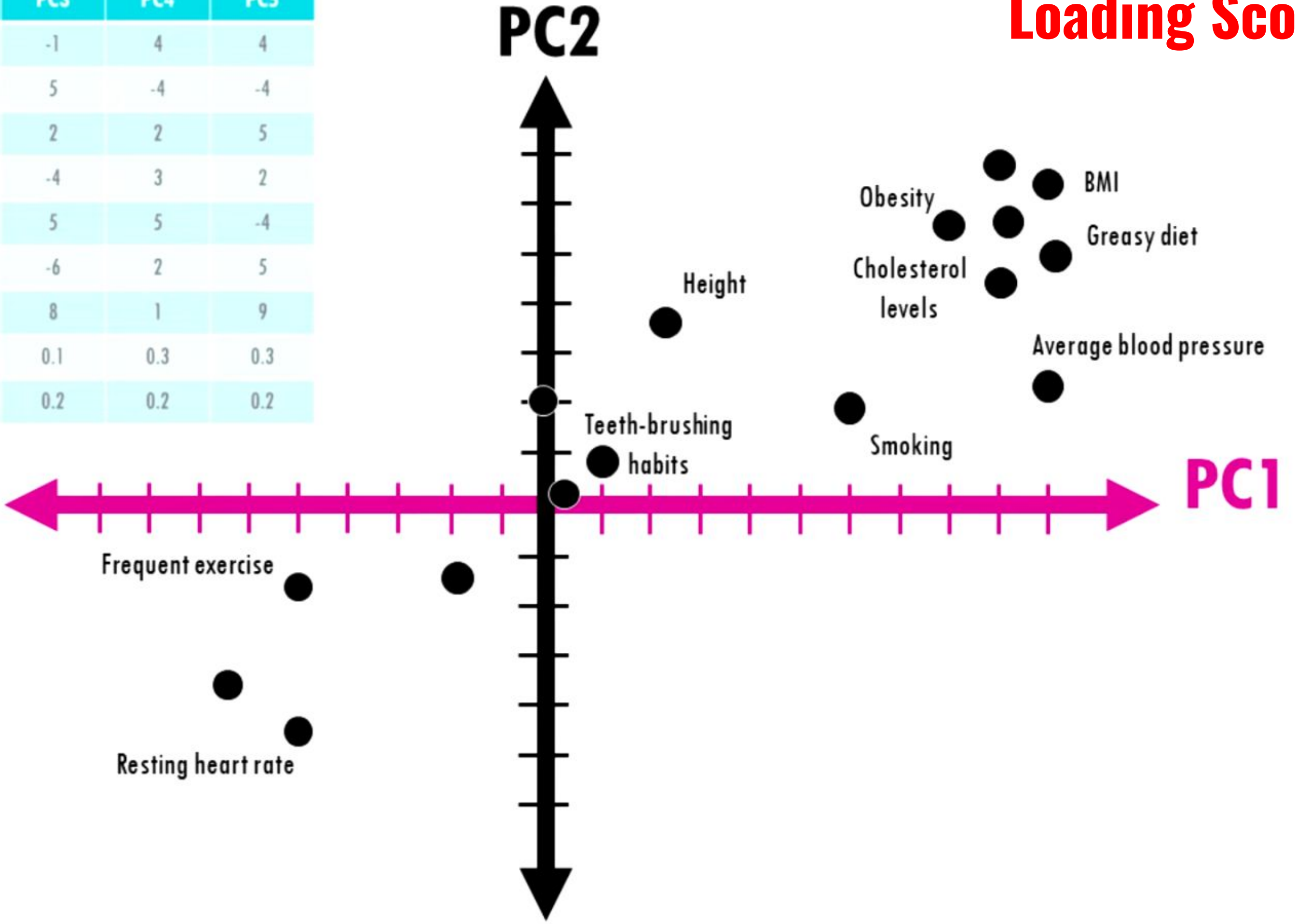


	PC1	PC2	PC3	PC4	PC5
Height	-1	3	-1	4	4
Average heart rate	9	7	5	-4	-4
BMI	10	6.5	2	2	5
Cholesterol levels	9	5	-4	3	2
Average cigarettes/day	7	2	5	5	-4
Greasy diet	10	5	-6	2	5
Frequent exercise	-5	-6	8	1	9
Eye colour	0.1	0.3	0.1	0.3	0.3
Teeth-brushing habits	0.2	0.2	0.2	0.2	0.2

Loading Score



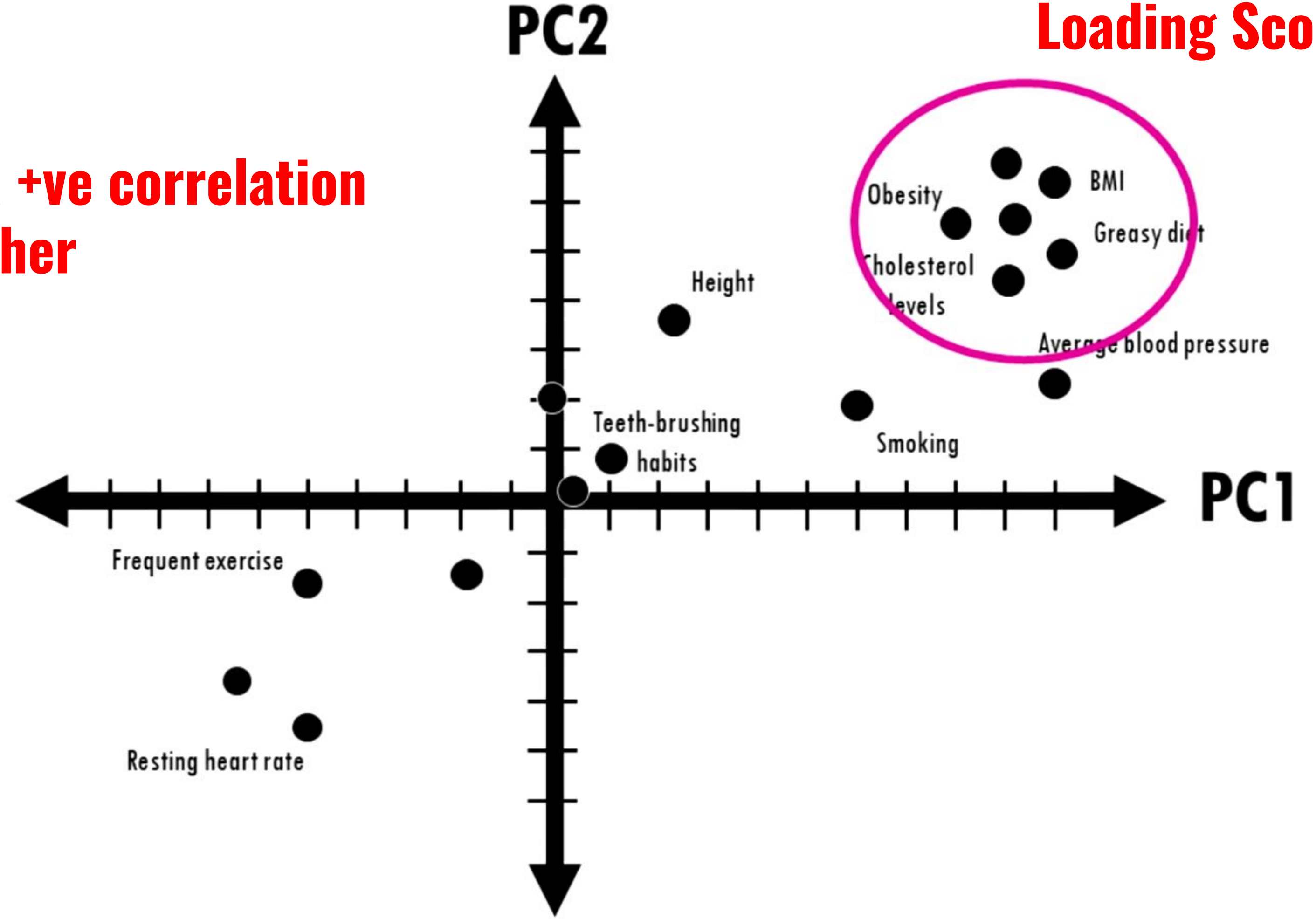
	PC1	PC2	PC3	PC4	PC5
Height	-1	3	-1	4	4
Average heart rate	9	7	5	-4	-4
BMI	10	6.5	2	2	5
Cholesterol levels	9	5	-4	3	2
Average cigarettes/day	7	2	5	5	-4
Greasy diet	10	5	-6	2	5
Frequent exercise	-5	-6	8	1	9
Eye colour	0.1	0.3	0.1	0.3	0.3
Teeth-brushing habits	0.2	0.2	0.2	0.2	0.2



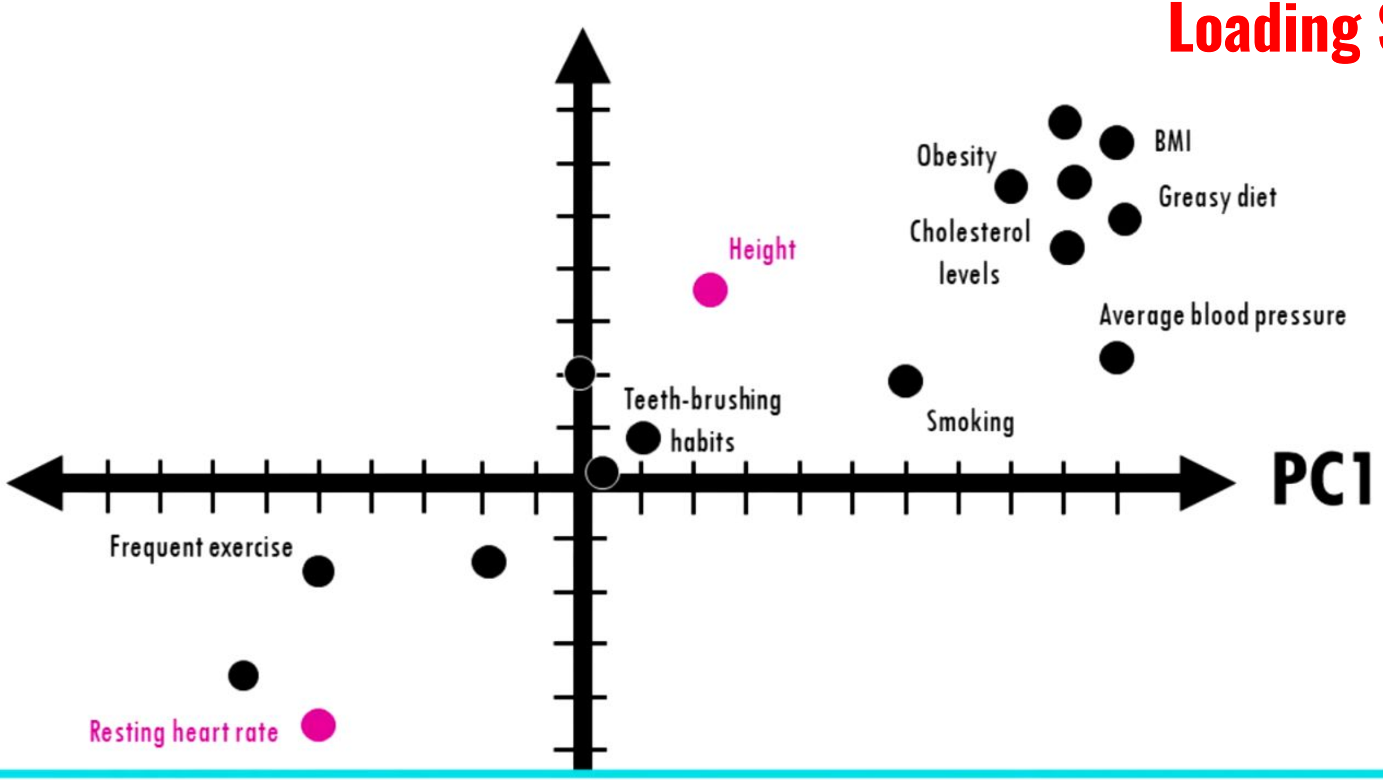
Loading Score:

Variables with a +ve correlation are group together

Loading Score



Loading Score

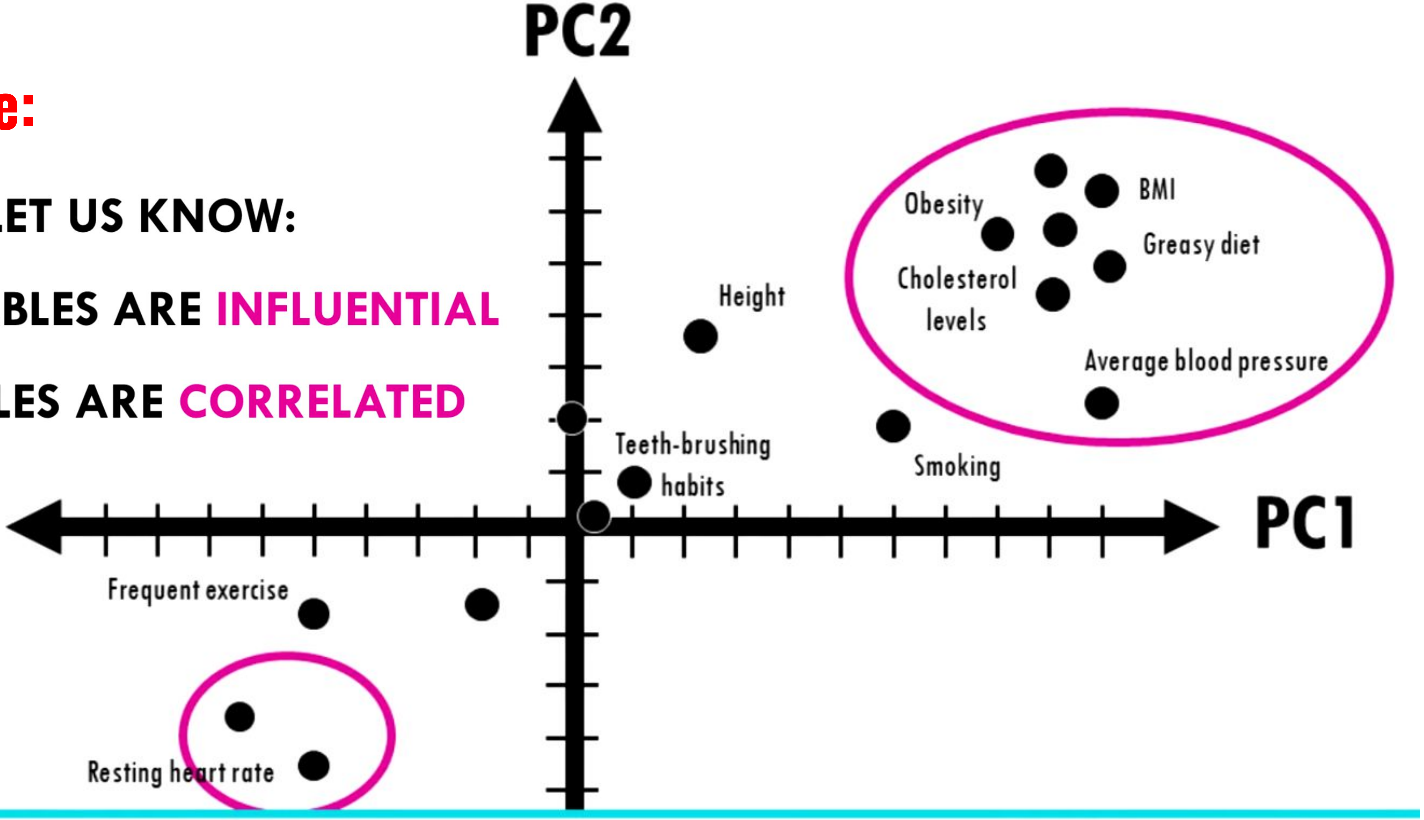


VARIABLES WITH A NEGATIVE CORRELATION ARE IN OPPOSITE SIDES OF THE ORIGIN

Loading Score:

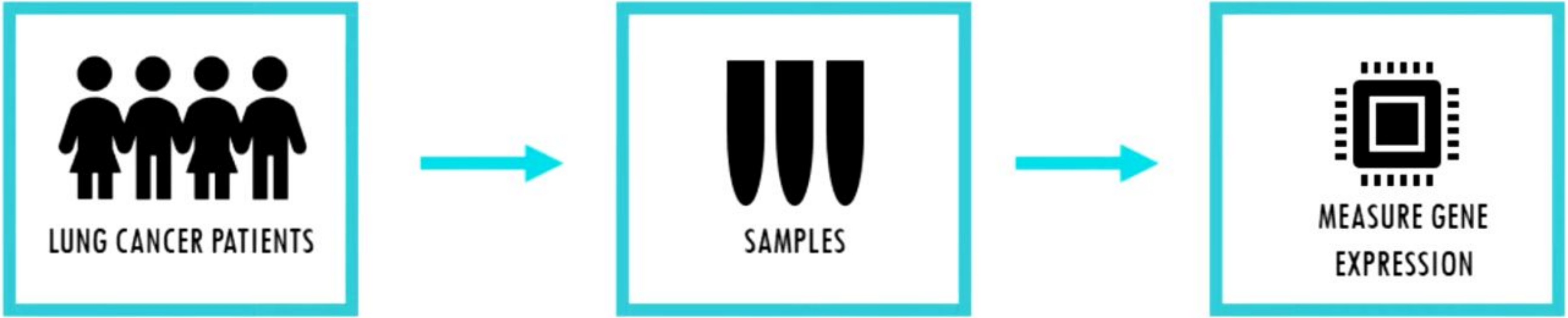
PC LOADINGS LET US KNOW:

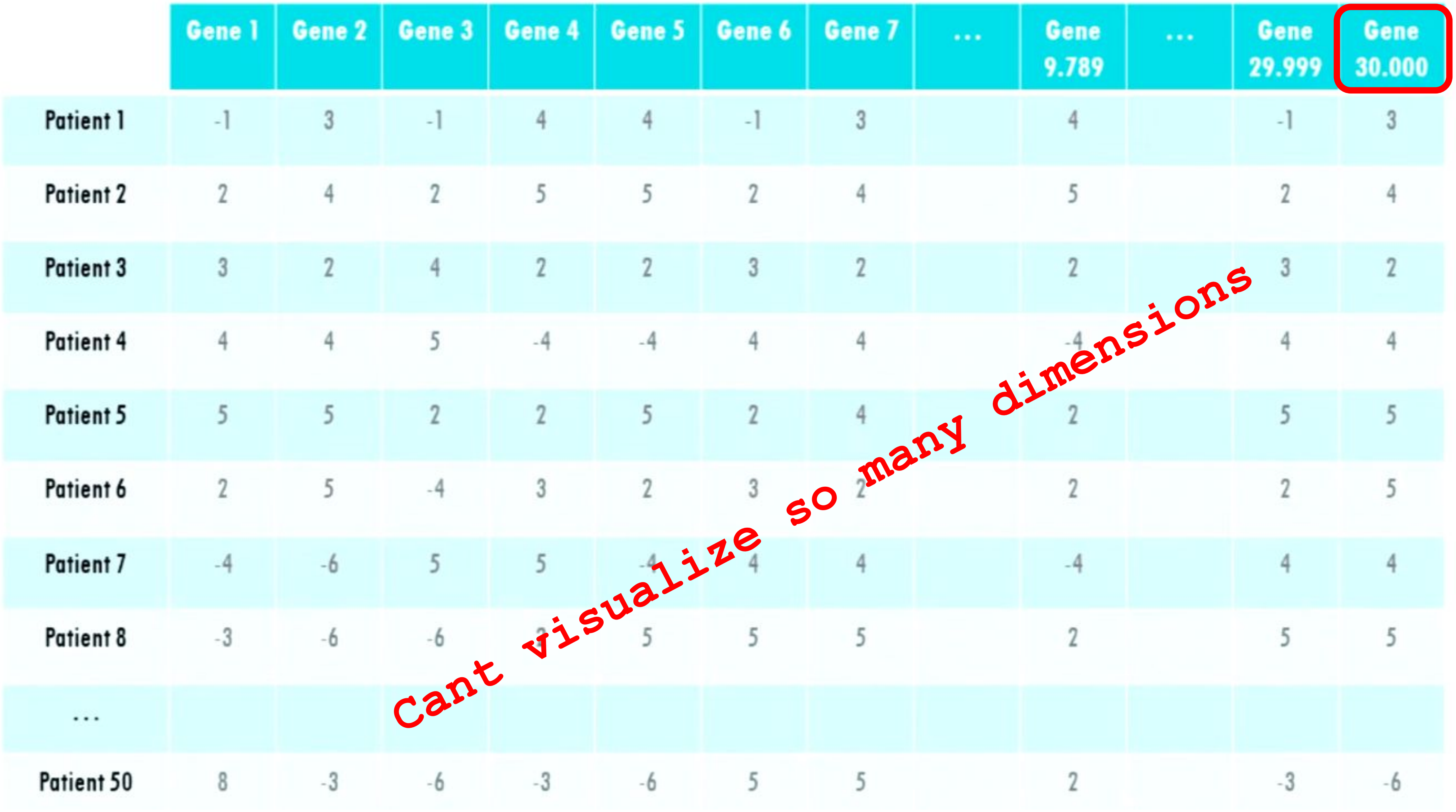
- WHICH VARIABLES ARE INFLUENTIAL
- HOW VARIABLES ARE CORRELATED



THE DISTANCE TO THE ORIGIN ALSO MATTERS!
LARGER WEIGHTS = BIGGER IMPACT

Example-2: Measure Gene Expression

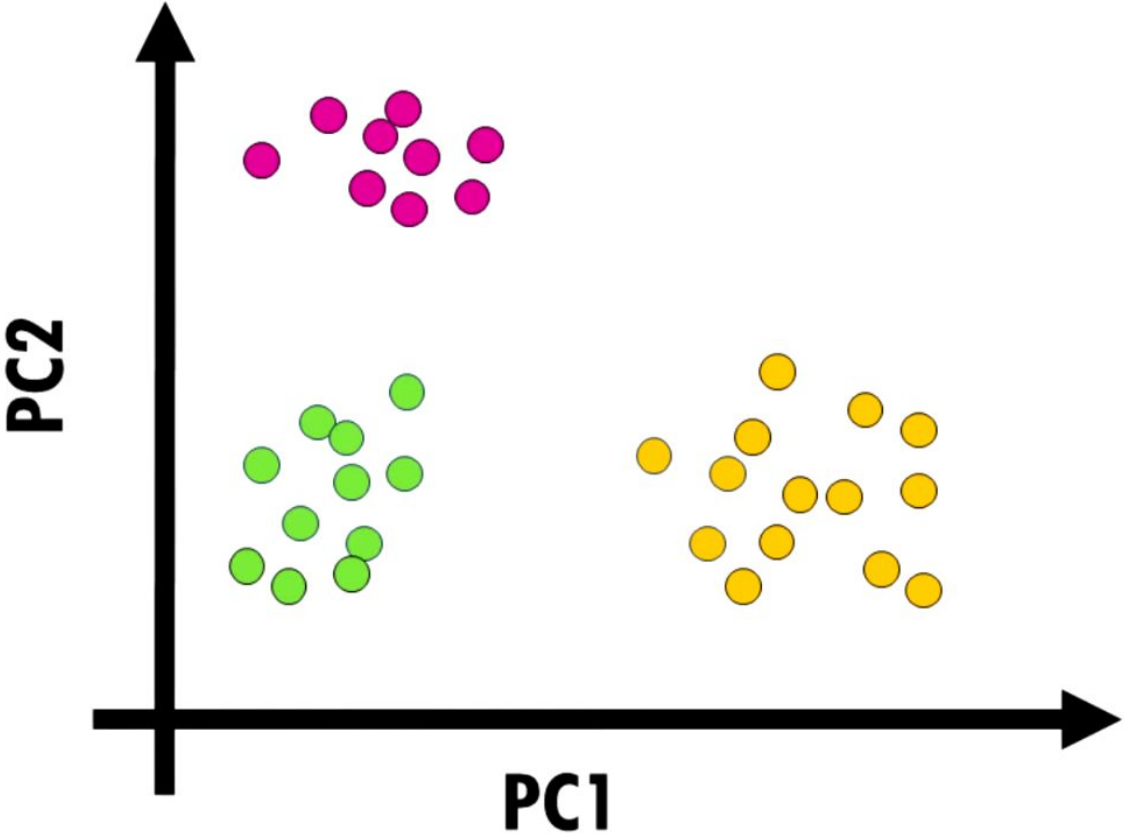




	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7	...	Gene 9.789	...	Gene 29.999	Gene 30.000
Patient 1	-1	3	-1	4	4	-1	3		4		-1	3
Patient 2	2	4	2	5	5	2	4		5		2	4
Patient 3	3	2	4	2	2	3	2		2		3	2
Patient 4	4	4	5	-4	-4	4	4		-4		4	4
Patient 5	5	5	2	2	5	2	4		2		5	5
Patient 6	2	5	-4	3	2	3	2		2		2	5
Patient 7	-4	-6	5	5	-4	4	4		-4		4	4
Patient 8	-3	-6	-6	2	5	5	5		2		5	5
...												
Patient 50	8	-3	-6	-3	-6	5	5		2		-3	-6

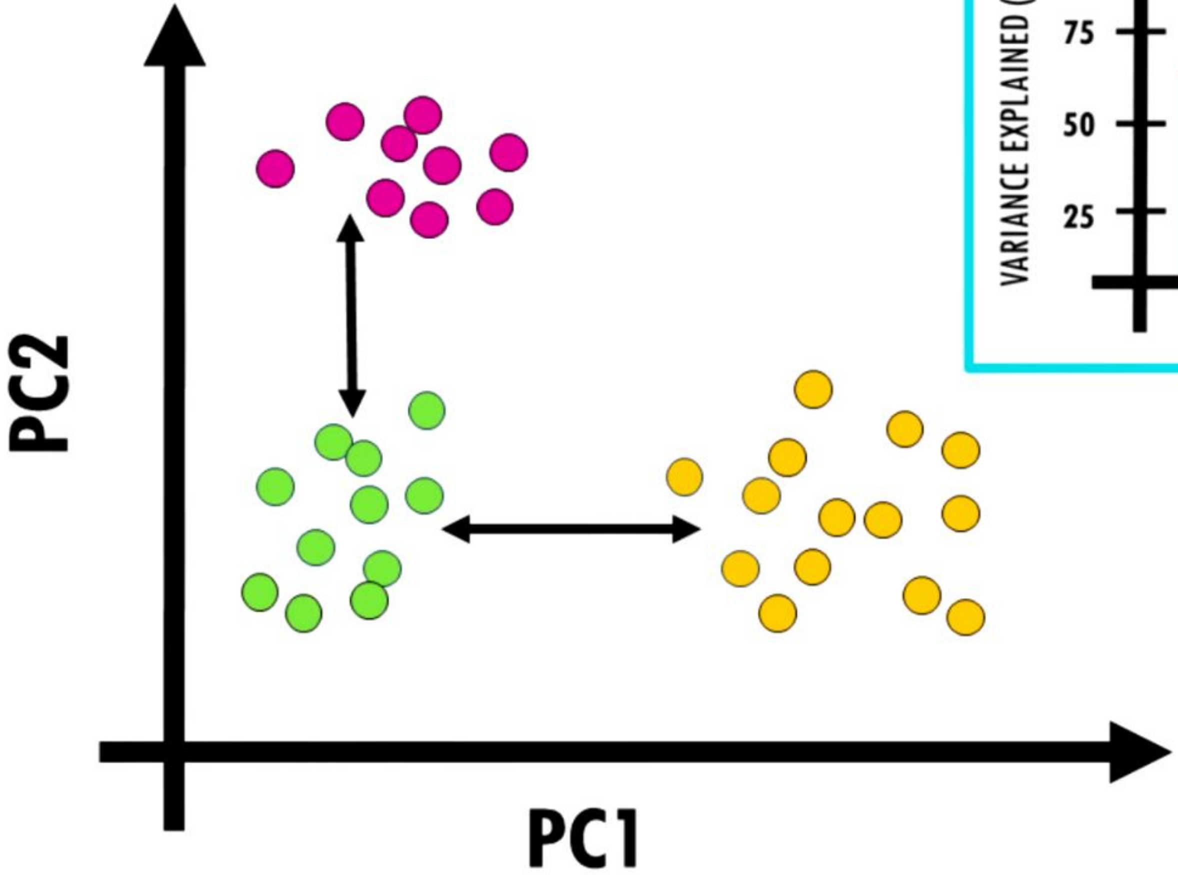
Cant visualize so many dimensions

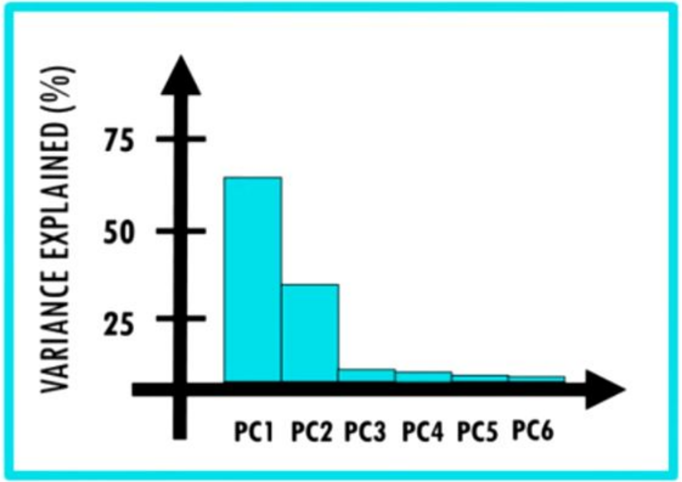
Example-2: Measure Gene Expression



Forms 3 clusters:
Reasons could be !!

- Drug-A
- Drug-B
- Radiotherapy





- PCA are ranked
- G & Y clusters are more different than G & P

Summary

- PCA summarises many dimensions into less (usually 2-3) by retaining as much information as possible.
- The SCREE Plot indicates how much variance (information) each PC holds
- Use PCA to visualise Trends, Jumps, Clusters, Outliers
  - Observations with similar overall profiles (PCA) are clustered together
  - Clusters separated by PC1 are more different than clusters separated by PC2