# Principal Component Analysis (PCA)
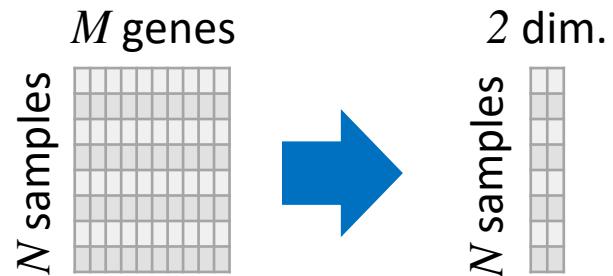
RNA-seq data analysis

**Paulo Czarnewski** | 13-May-2019

# Why PCA?

Simplify complexity, so it becomes easier to work with.
- Reduce number of features (genes)
- Need to transform non-linear relationships to linear
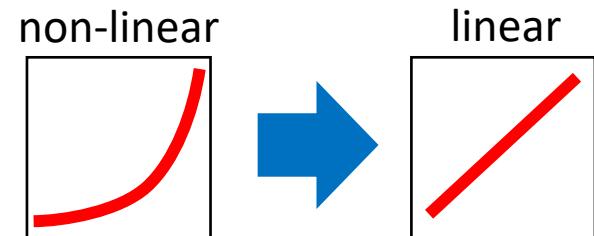

*M* genes → *2* dim.
*N* samples

"Remove" redundancies in the data

Identify the most relevant information
*Find and filter noise*


non-linear → linear
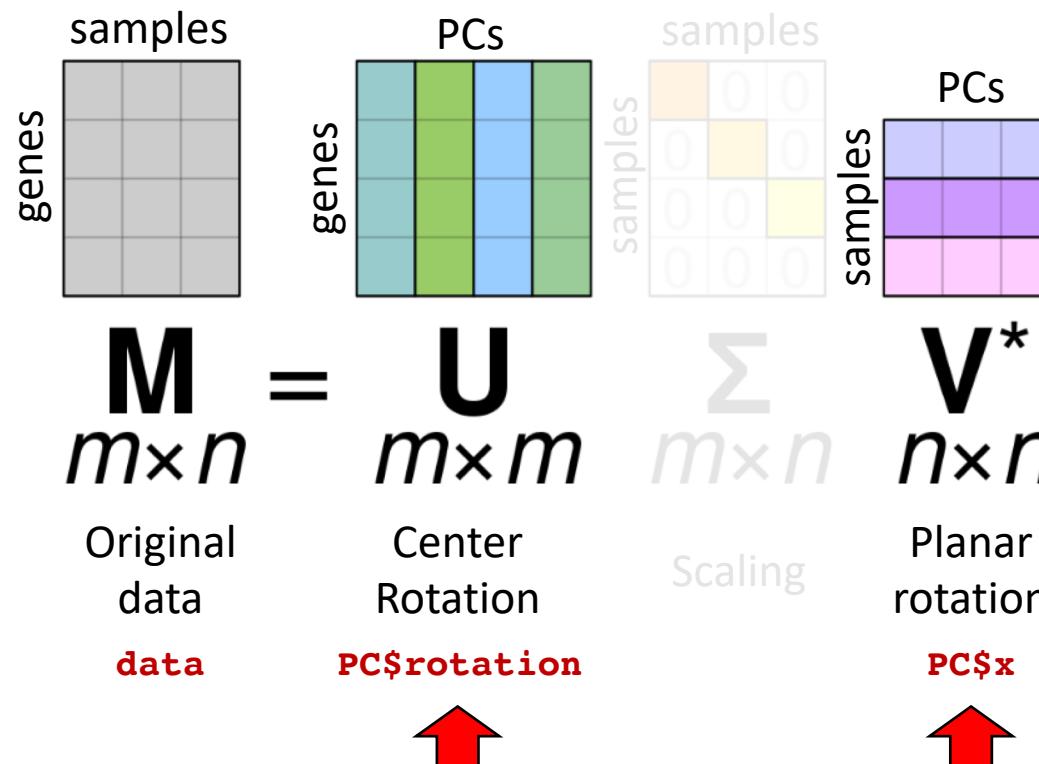
Data visualization

# How PCA works

It is a <u>LINEAR</u> algebraic method of dimensionality reduction.

It is a case inside Singular Value Decomposition (SVD) method (data compression)
*Any matrix can be decomposed as a multiplication of other matrices (Matrix Factorization).*



$$\underset{m \times n}{\mathbf{M}} = \underset{m \times m}{\mathbf{U}} \quad \underset{m \times n}{\Sigma} \quad \underset{n \times n}{\mathbf{V}^*}$$

|  |  |  |  |
|---|---|---|---|
| Original data | Center Rotation | Scaling | Planar rotation |
| **data** | **PC$rotation** |  | **PC$x** |
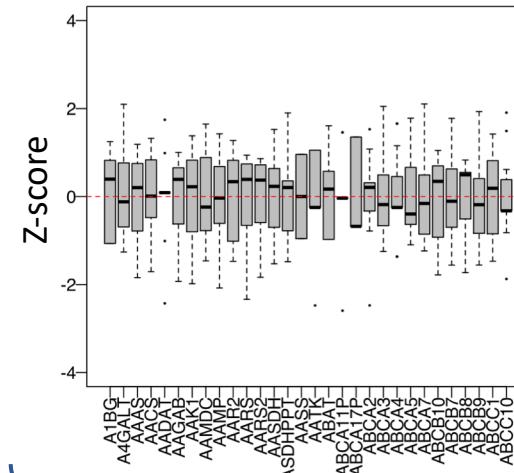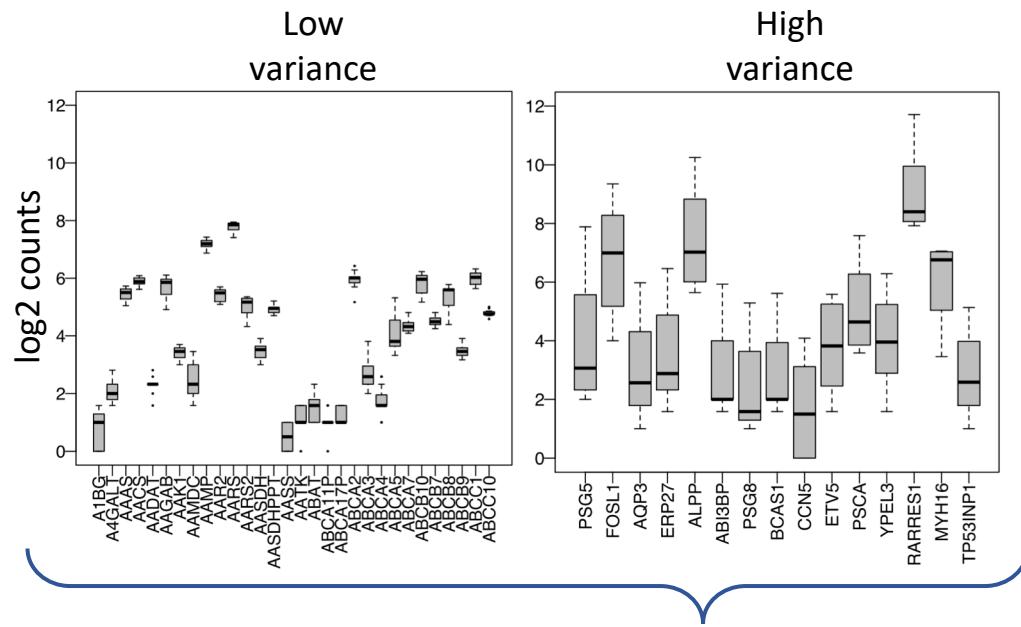
```
PC <- prcomp( data )
PC$|

   ◇ sdev
   ▦ rotation
   ◆ center
   ◆ scale
   ▦ x
```

# Data transformation and scaling

Before applying PCA, the data should be first transformed to a <u>linear</u> scale (i.e. log)

Each feature should be scaled to have a similar center (zero) and similar deviation.
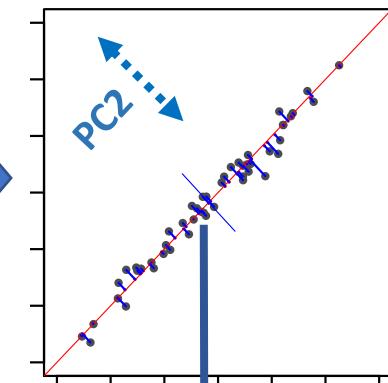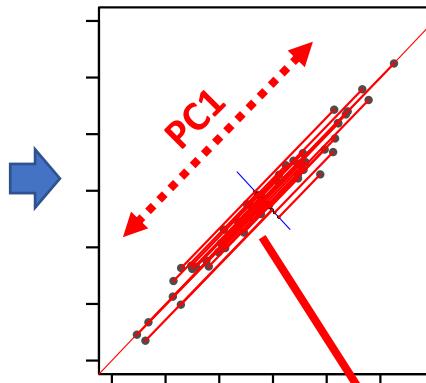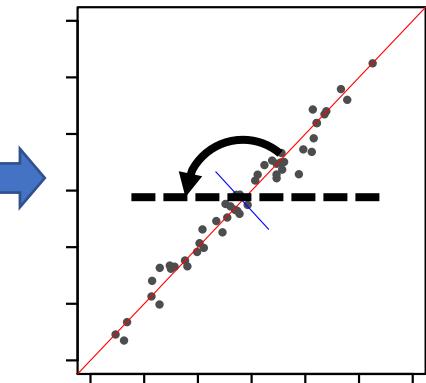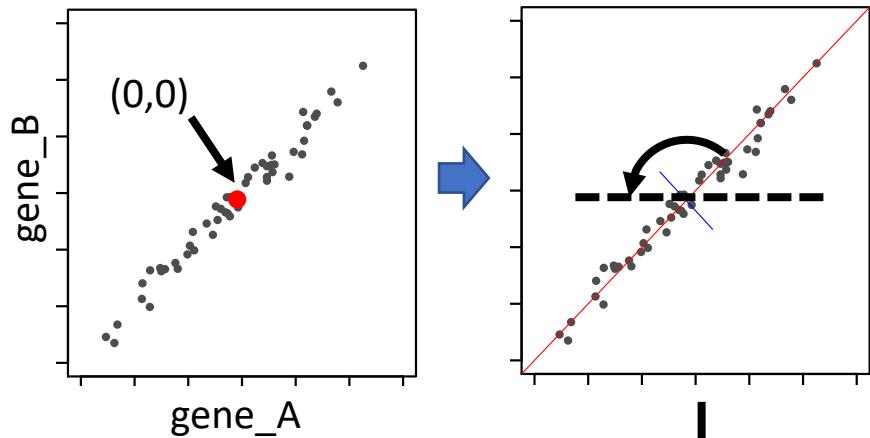


PCA on raw counts will separate genes with <u>higher counts</u> in the first PCs

*(higher distance to 0)*
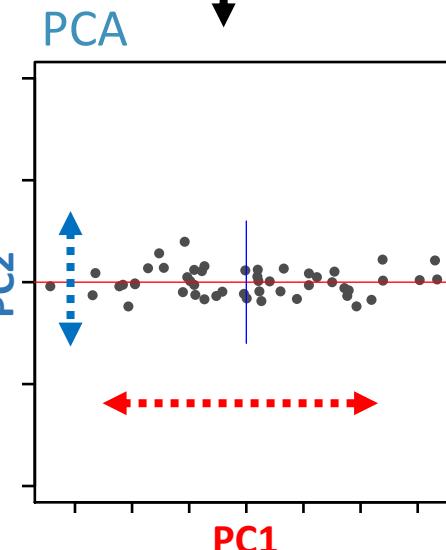
PCA on Z-score will separate genes with most <u>common expression trends</u> in the first PCs

$$M = U \Sigma V^*$$
$$m \times n \quad m \times m \quad m \times n \quad n \times n$$

scaling

# How PCA works



original data (Z-score)
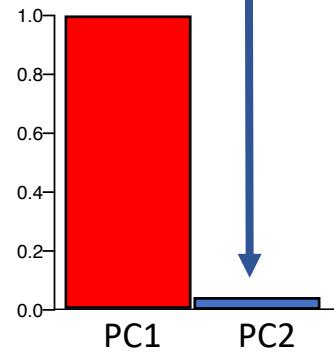
rotation to another coordinate system

PCA

percentage of variance explained

$$\underset{m\times n}{\mathbf{M}} = \underset{m\times m}{\mathbf{U}} \quad \underset{m\times n}{\mathbf{\Sigma}} \quad \underset{n\times n}{\mathbf{V}}$$
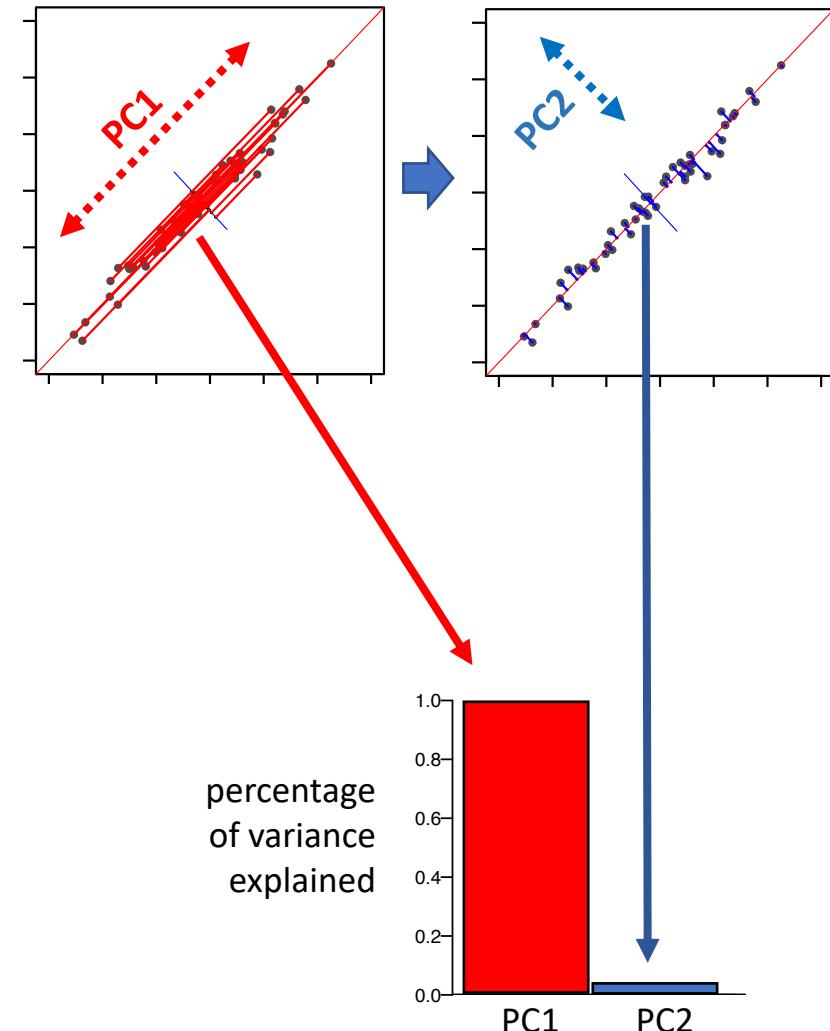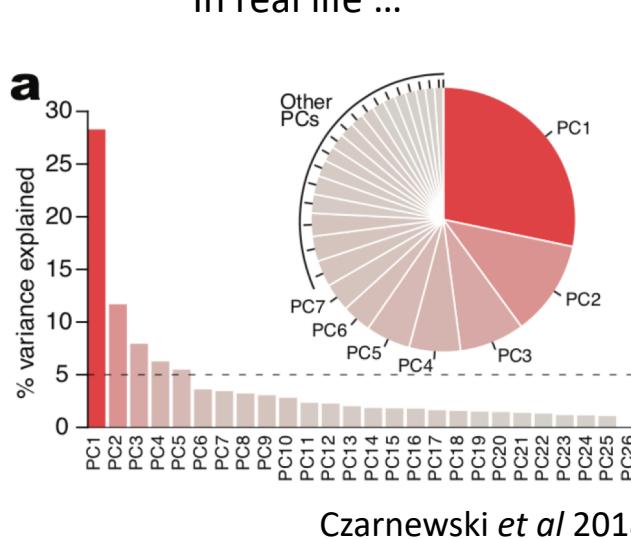
rotation

# How PCA works

PC1 explains >98% of the variance
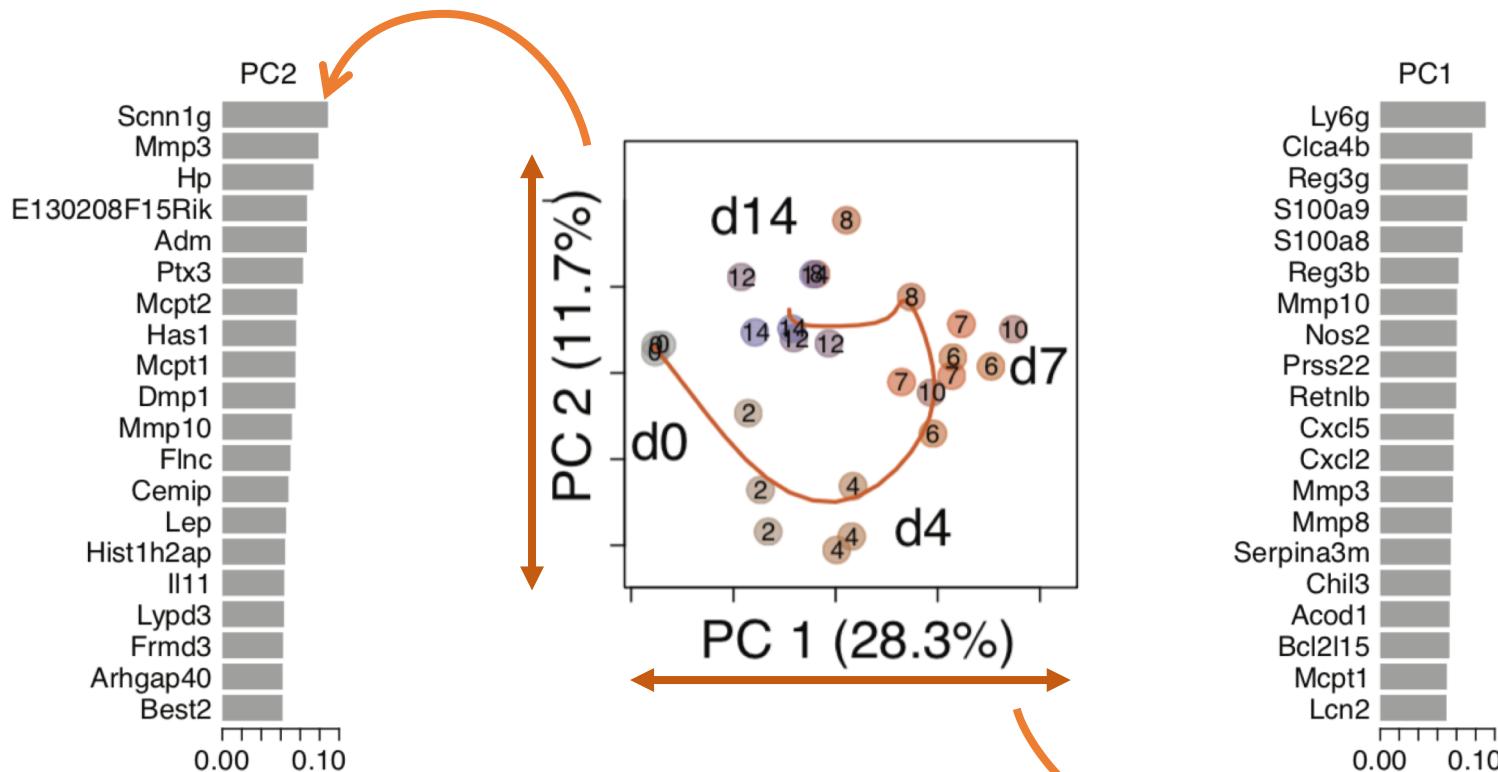
1 PC thus represents 2 genes very well
*"Removing" redundancy*

PC2 is nearly insignificant in this example
*Could be disregarded*

In real life …



Czarnewski *et al* 2018



percentage
of variance
explained

# How PCA works

Each PC has a meaning



Czarnewski *et al* 2018

# A visual intuition of PCA

The top principal components store more important ~~Shakira~~ information

Original



PCA → $V^*$ $n{\times}n$ → Use only # PCs

1 PC | 10 PCs | 20 PCs
30 PCs | 40 PCs | 50 PCs

# PCA summary
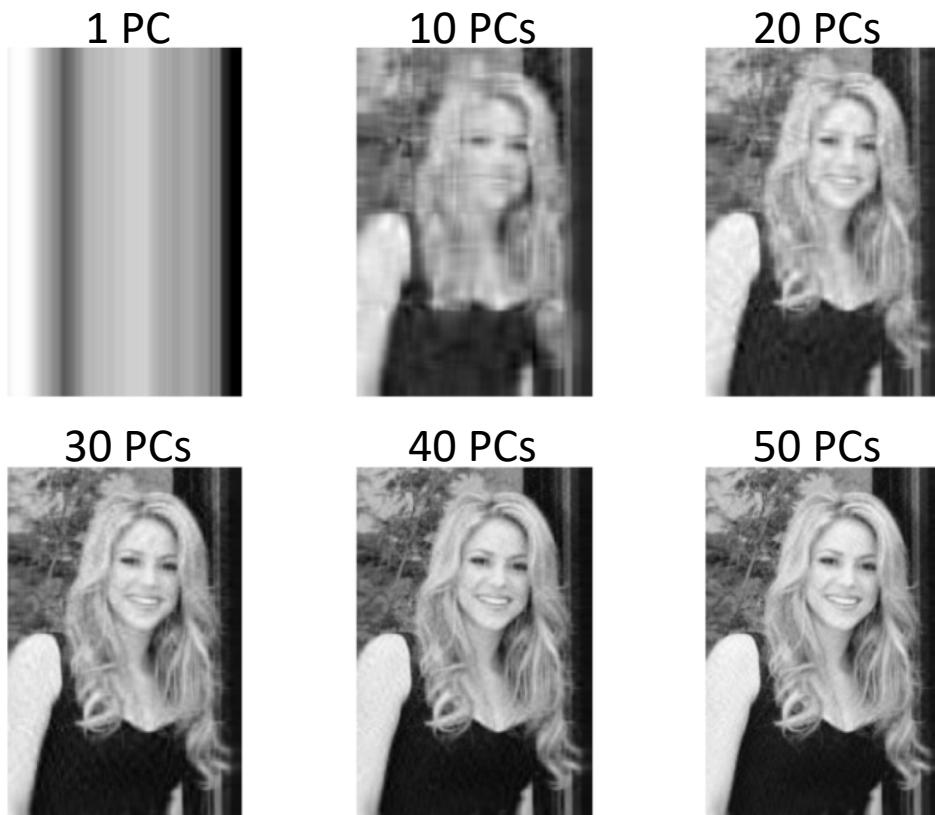
It is a <u>LINEAR</u> method of dimensionality reduction
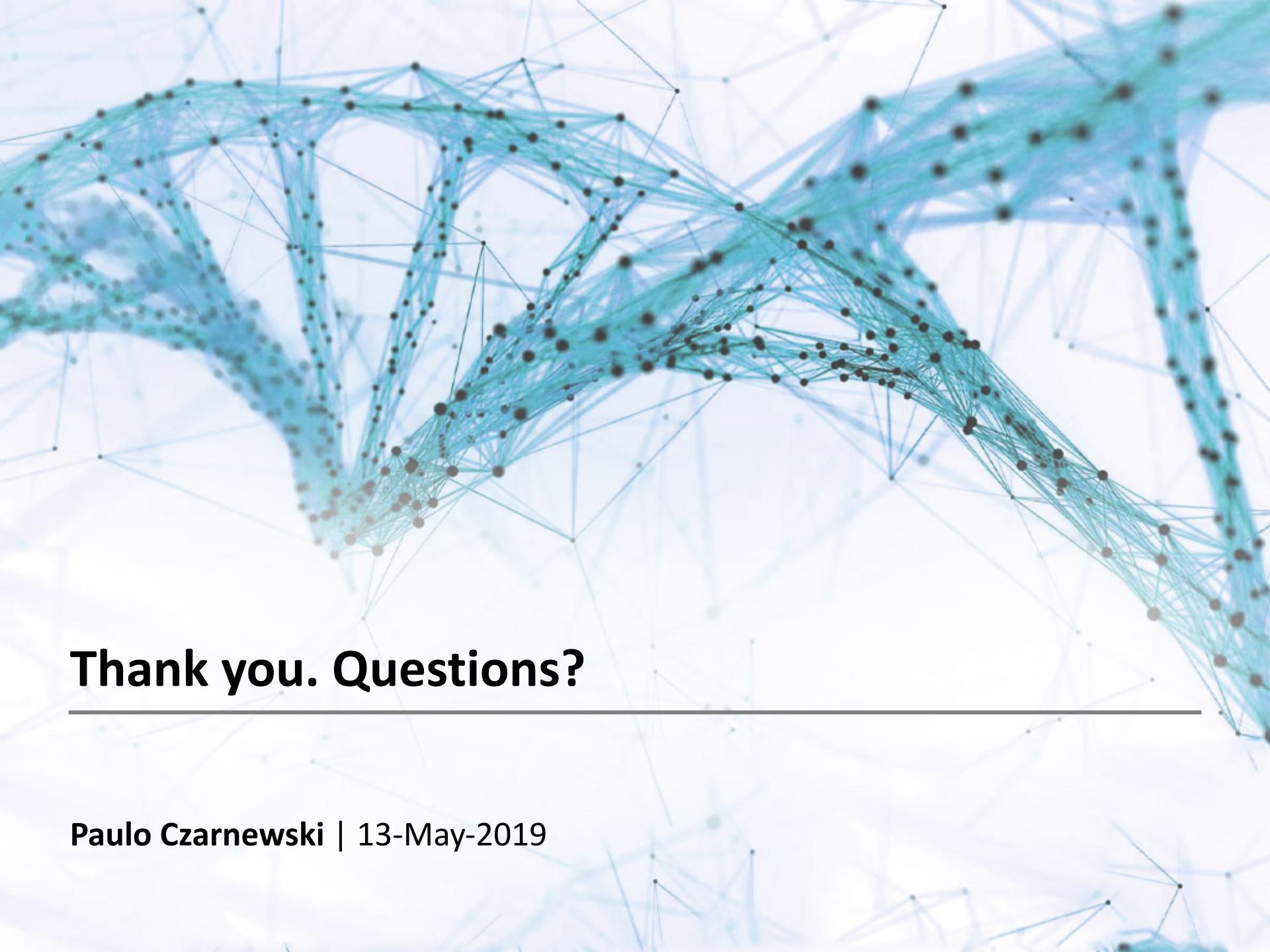
The data is usually <u>SCALED</u> (i.e. Z-score ) and TRANSFORMED (i.e. log) prior to PCA

It is an <u>interpretable</u> dimensionality reduction

The top principal components contain <u>higher variance </u>from the data

Can be used as <u>FILTERING</u>, by selecting only the top significant PCs
- PCs that explain at least 1% of variance
- The first 5-10 PCs

# Thank you. Questions?

**Paulo Czarnewski** | 13-May-2019