Processing 1

Normalization, scaling and regression

Bastien Job, Gustave Roussy, Villejuif

Nathalie Lehmann, Institut Pasteur

EBAII 2024 - 20.11.2024

Organisation of the scRNA-seq course

- From raw count matrix to normalised matrix
  - Filtering low quality droplets
  - Filtering dead cells
  - Filtering doublets
- Data normalization
  - Why do we need to normalize the data ?
  - What are the methods available ?
  - Regression of biological biaises

Question !

What do you think of each gene expression in this matrix ?

Gène	Cellule 1	Cellule 2	Cellule 3	Cellule 4
Rouge	100	200	300	400
Bleu	50	100	150	200
Vert	10	10	10	10
Jaune	100	100	100	100

Why do we need to normalize our data ?

We need to remove **technical biases** in order to…

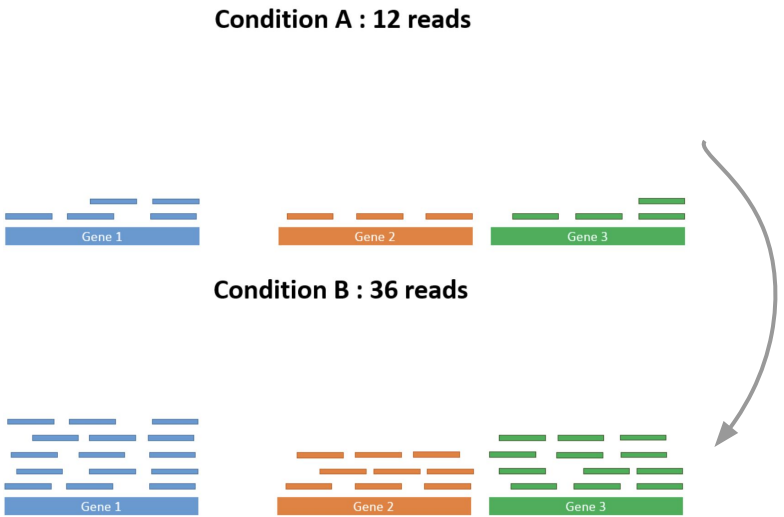
Condition A : 12 reads

Condition B : 36 reads

Images from https://www.biostars.org/p/349881/

Why do we need to normalize our data ?

We need to remove **technical biases** in order to…



The 2 libraries have the **same RNA composition**.

But the condition B has 3 times more reads than the condition A.

We need to correct for differences in **library size**.

Images from https://www.biostars.org/p/349881/

Why do we need to normalize our data ?

We need to remove **technical biases** in order to…

Normalization allows us to compare cells

Question !

What do you know about normalization ?

How many approaches do you think there are ?

Plenty of normalization approaches for bulk RNA-seq

- TPM
- CPM

- RPKM
- FPKM

- Global scaling (eg: Upper Quartile)

- Size factors calculation (eg: estimation of library sampling depth) :
  - DESeq2
  - edgeR

- ...

Plenty of normalization approaches for bulk RNA-seq

- TPM
- CPM

- RPKM
- FPKM

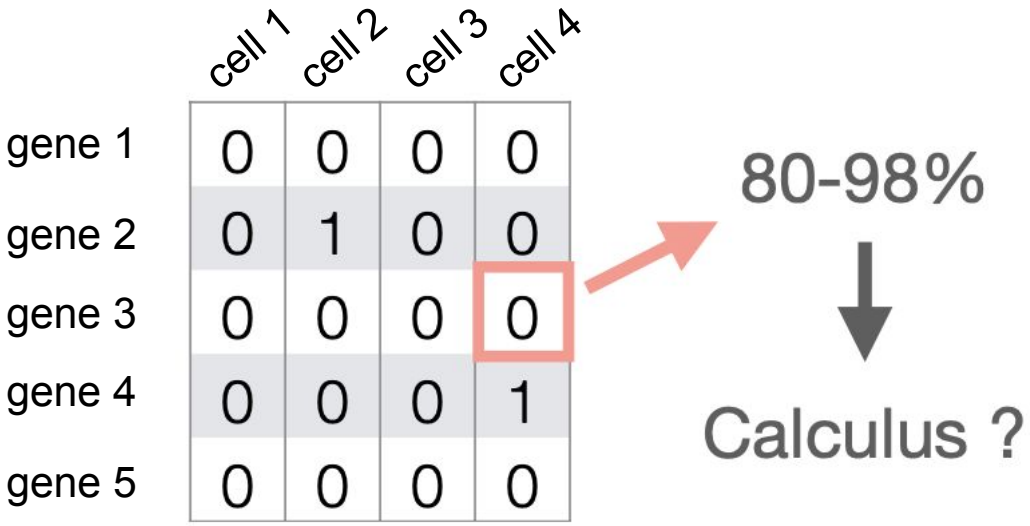
These methods do not apply to single-cell data (or partially)

- Global scaling (eg: Upper Quartile)

- Size factors calculation (eg: estimation of library sampling depth) :
  - DESeq2
  - edgeR

- ...

This is mostly due to the sparsity of the single-cell data



A **sparse matrix** is a matrix filled with a LOT of zeros

How do we normalize data in basic Seurat ?

First, let’s take this simple raw count matrix :

Gène	Cellule 1	Cellule 2	Cellule 3	Cellule 4
Rouge	100	200	300	400
Bleu	50	100	150	200
Vert	10	10	10	10
Jaune	100	100	100	100

How do we normalize data in basic Seurat ?

First, let's take this simple raw count matrix :

Gène	Cellule 1	Cellule 2	Cellule 3	Cellule 4
Rouge	100	200	300	400
Bleu	50	100	150	200
Vert	10	10	10	10
Jaune	100	100	100	100

Our first impressions are like :

Gene rouge : expression is proportional to the sequencing depth
Gene bleu : same as "gene rouge" but with smaller values
Gene vert : expression is low and steady
Gene jaune : expression is steady between the cells

How do we normalize data in basic Seurat ?

First, let's take this simple raw count matrix :

Gène	Cellule 1	Cellule 2	Cellule 3	Cellule 4
Rouge	100	200	300	400
Bleu	50	100	150	200
Vert	10	10	10	10
Jaune	100	100	100	100

Sequencing depth : **260** 410 560 710

CPM normalized value for gene rouge in cellule 1 : 100/**260*****10**^{4}= 6250

How do we normalize data in basic Seurat ?

Let's apply this to all genes :

Gène	Cellule 1	Cellule 2	Cellule 3	Cellule 4
Rouge	6250	6451	6522	6557
Bleu	3125	3226	3261	3279
Vert	625	323	217	164
Jaune	6250	3226	2174	1639

After CPM normalization, we have :

Gene rouge : the expression is actually steady between the cells
Gene bleu : the expression is actually steady between the cells (but lower than gene rouge)
Gene vert : expression decreases
Gene jaune : expression decreases

How do we normalize data in basic Seurat ?

We applied this formula (we just omitted the log and the pseudo count “+1”) :

$$\text{Normalized value}_{g,c} = \log_2 \left(\frac{\text{UMI}_{g,c}}{\text{TotalUMI}_c} \times \text{ScaleFactor} + 1 \right)$$

Why do we add the log transformation ?

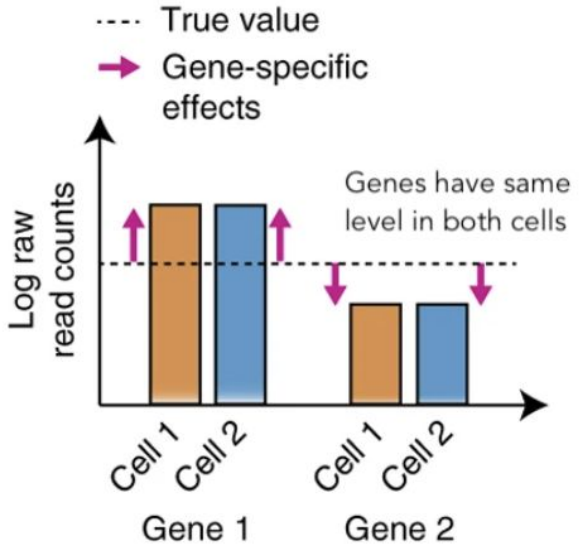
- To reduce the impact of extreme values
- To stabilize the variance
- To have a better biological interpretation
- Improved Model Fit

Question !

What about scaling ???

Why do we scale ?

- To be able to improve comparability between **genes**



Examples of biological biases that you may want to correct :

- Amplification
- RNA capture efficiency
- Gene length
- GC content

Image adapted from *Vallejos et al., 2017*

How do we scale ?

Let’s scale gene rouge :

Gène	Cellule 1	Cellule 2	Cellule 3	Cellule 4
Rouge	6250	6451	6522	6557
Bleu	3125	3226	3261	3279
Vert	625	323	217	164
Jaune	6250	3226	2174	1639

Mean expression of gene rouge = 25780 / 4 = 6445
Standard deviation $\sigma \approx 137.3$

How do we scale ?

Let’s scale gene rouge :

Mean expression of gene rouge = 25780 / 4 = 6445
Standard deviation σ ≈ 137.3

Use the formula:

$$\text{Scaled\_value}_i = \frac{\text{Value}_i - \mu}{\sigma}$$

Calculations for each cell:

$$\text{For } 6250: \quad \frac{6250 - 6445}{137.3} = \frac{-195}{137.3} \approx -1.42$$

$$\text{For } 6451: \quad \frac{6451 - 6445}{137.3} = \frac{6}{137.3} \approx 0.04$$

$$\text{For } 6522: \quad \frac{6522 - 6445}{137.3} = \frac{77}{137.3} \approx 0.56$$

$$\text{For } 6557: \quad \frac{6557 - 6445}{137.3} = \frac{112}{137.3} \approx 0.82$$

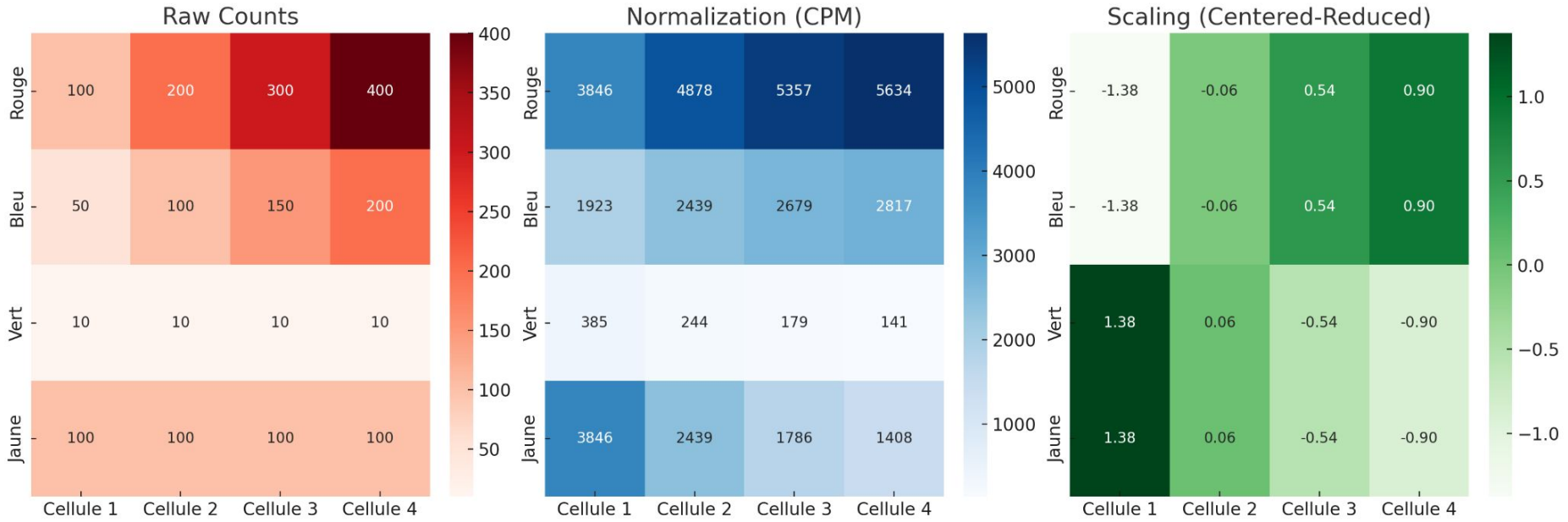
This is the z-score !

Key points of scaling

- Scaling ensures that the values have a mean of 0 and a standard deviation of 1

- This transformation makes genes with different ranges comparable across cells or datasets

Visualisation of the impact of each transformation



When to avoid scaling ?

Be careful ! Scaling may not always be necessary…

When to avoid scaling ?

Be careful ! Scaling may not always be necessary…

For example:

- In differential expression analysis, raw or log-normalized counts are used since absolute expression levels matter.
- For some downstream analyses like gene set enrichment analysis (GSEA), raw expression values are often more meaningful.

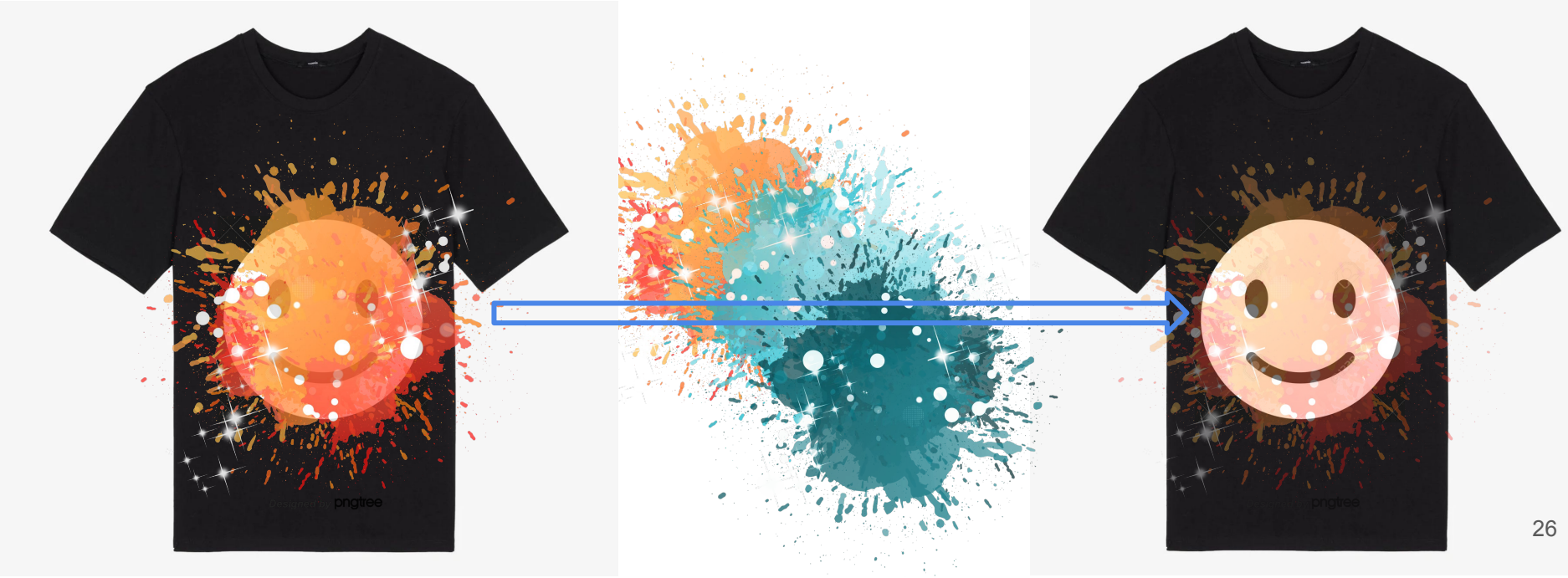
What about regression ?

HO NO !

- Some close relative offered you your favorite gear : a nice T-shirt with a cool design
- Unfortunately your 5-years old inadvertently tainted it with its paint flask…
- Hopefully, you know a skilled team of cleaners right around the corner (who actually are bioinformaticians in disguise ! **wink wink**)
- They explain you they learned how to clean stains by **accumulating knowledge**, having cleaning tons of them in several contexts.

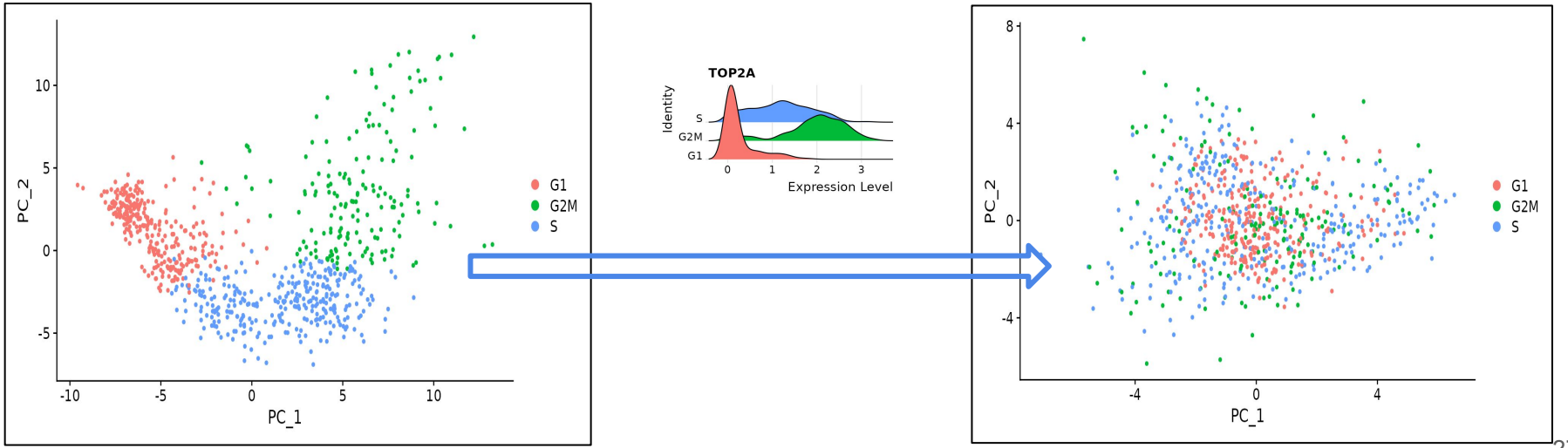
Cleaners to the rescue

- By compiling this amount of knowledge, they could produce a specific cleaning protocol to the kind of staining affecting your Tee, subtracted the former to the latter, revealing the object of desire



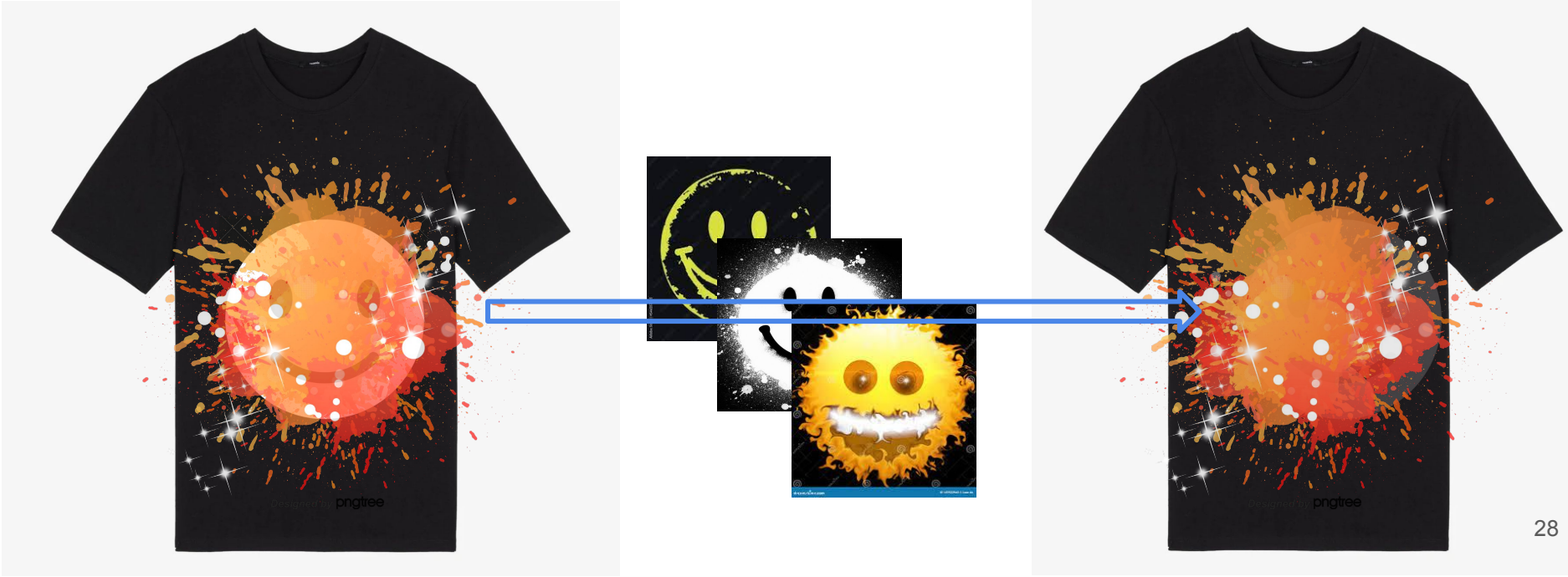
Let’s get back to reality

- This is what is performed in the cell cycle example, thanks to acquired data on the cell-cycle -specific gene expression, and regression

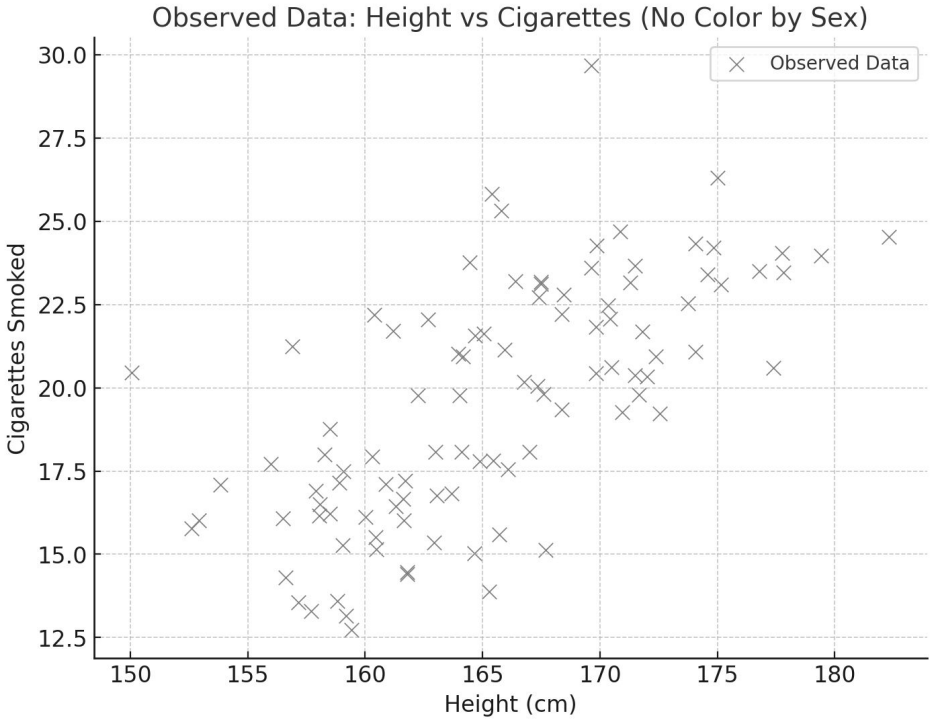


The devil may be in the bottle

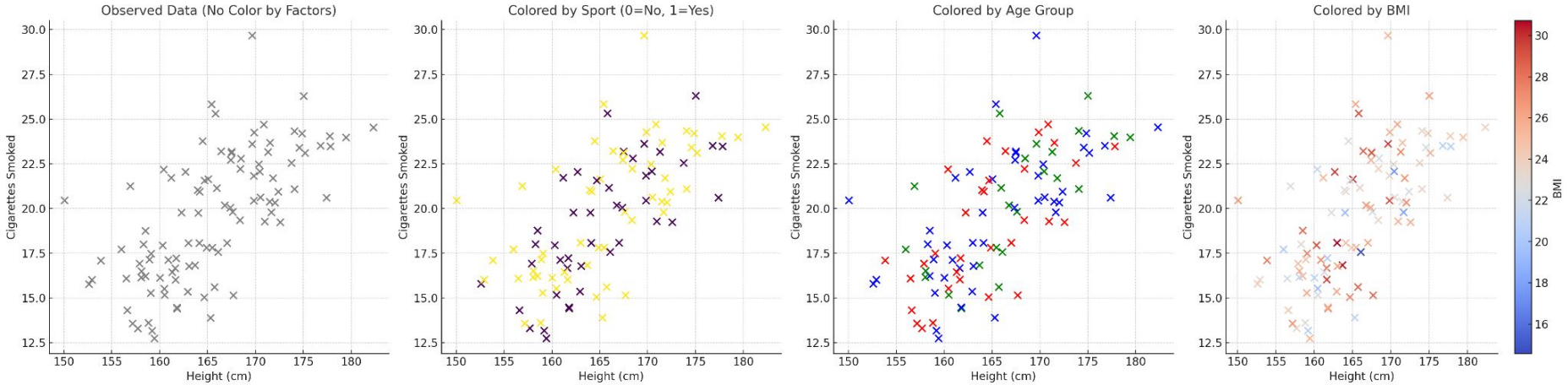
- But what would happen if by bad luck, the cleaner team trained themselves on stains that unfortunately ***look a bit to much like*** your pattern ?
- The same way you **increased signal by removing some noise** on the former example, one may perform the **complete opposite** if one doesn't pay attention
- *Probably because in this specific context, what you thought was noise, wasn't*



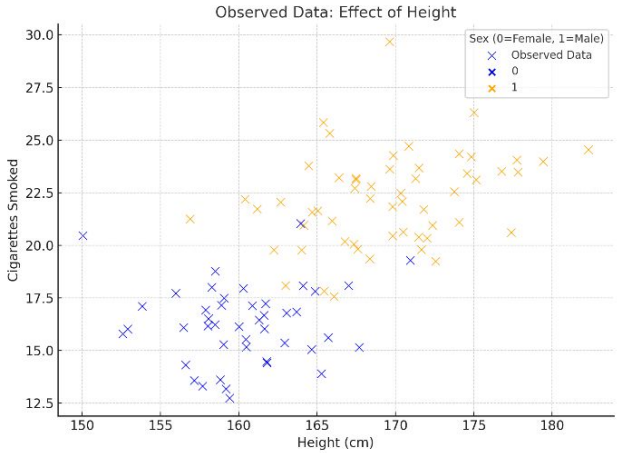
What about regression ?

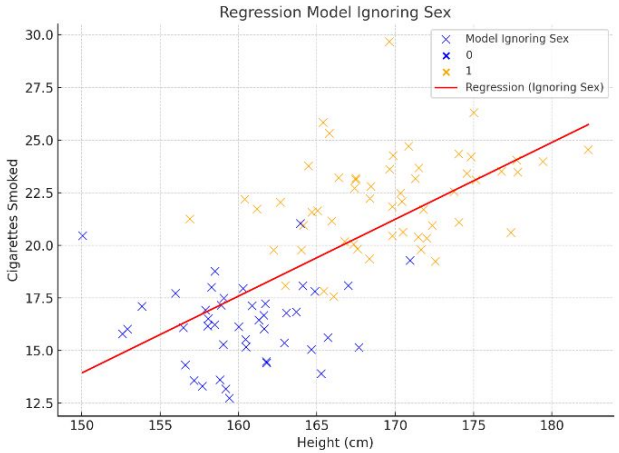


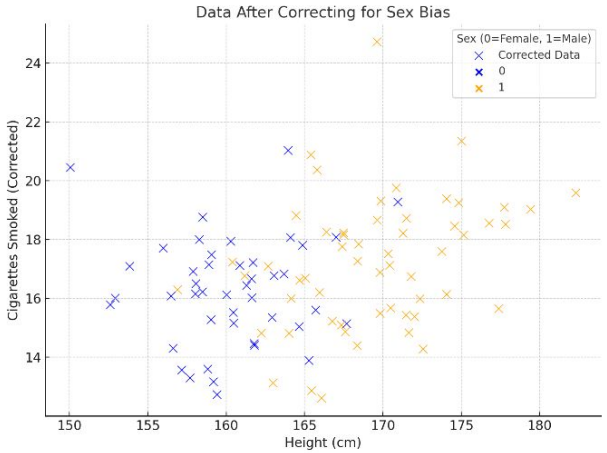
What about regression ?



What about regression ?







Acknowledgements

- Some illustrations/slide were created by Marine Aglave