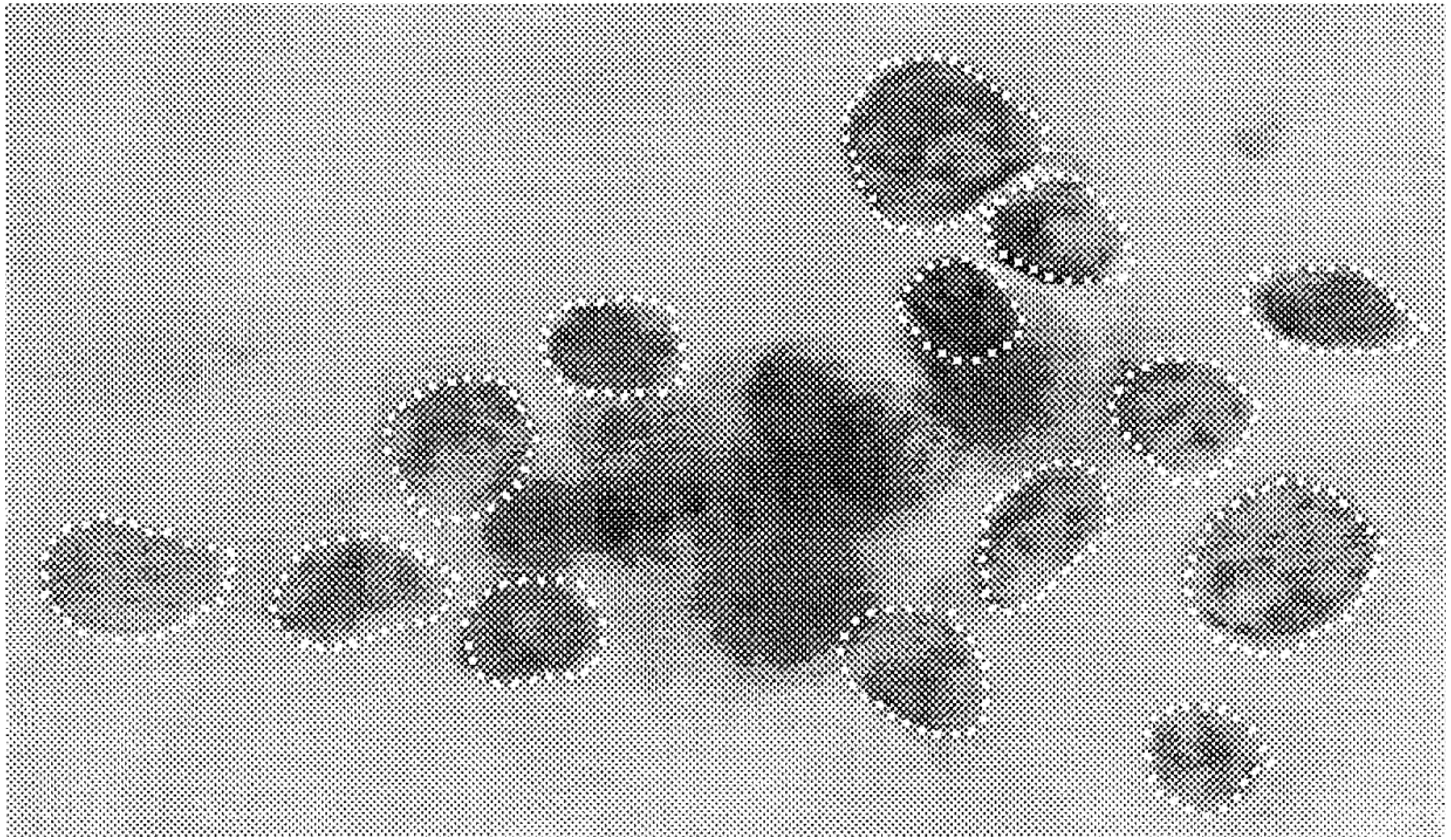
[Tuxae]



[Ateliers Data]/[2. Classification - Prédire le cancer du sein ?]

Introduction

Jupyter-python

The JupyterLab IDE with Python, Julia, and a collection of standard data science packages.

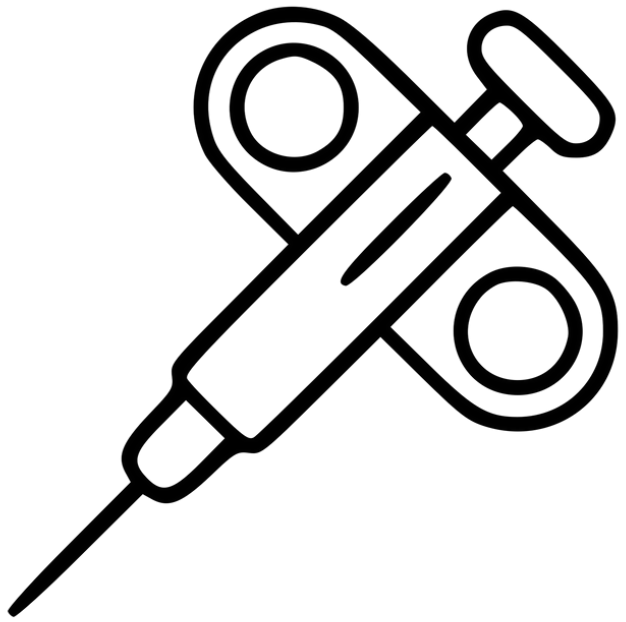
Learn more Launch

https://shorturl.at/EKbhK

Le problème

Le problème

1992

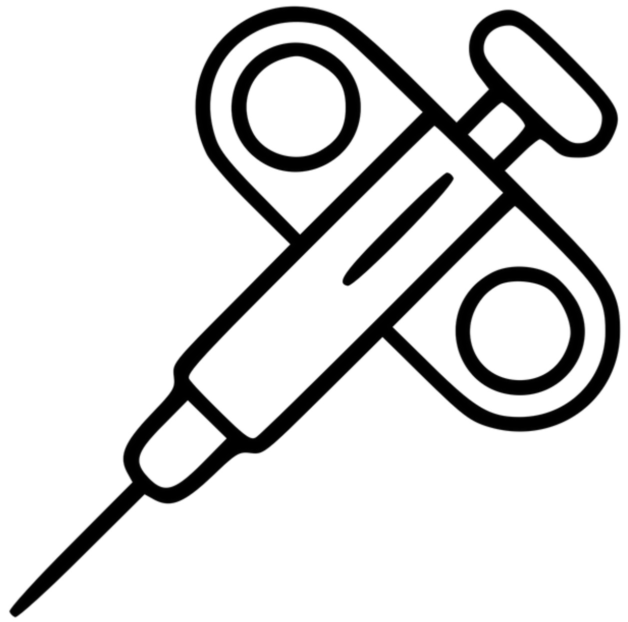


Biopsie totale

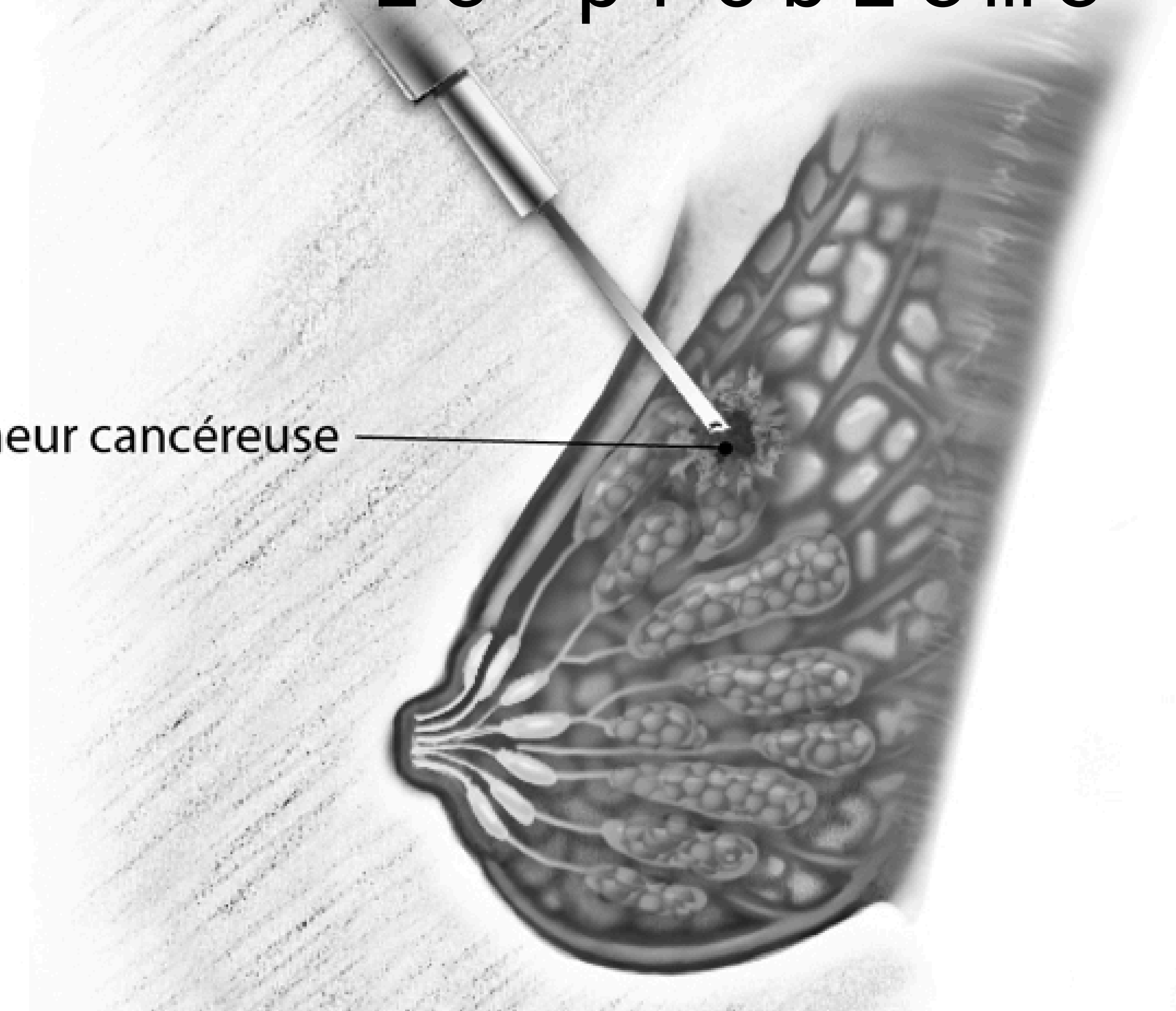
BIOPSIE

Le problème

1992



Biopsie totale



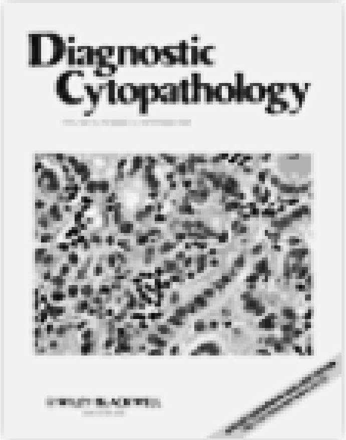
biopsie à l'aiguille fine

Le problème

FINE-NEEDLE ASPIRATION BIOPSY: A HISTORICAL OVERVIEW

Fine needle aspiration biopsy is a safe, inexpensive and accurate technique for the diagnosis of benign and malignant conditions. [...] However, **the situation was not always like this**. In its beginnings the procedure suffered from **all kinds of criticism and attacks**. [...]

Diagn. Cytopathol. 2008;36:773–775. © 2008 Wiley-Liss, Inc.



Volume 36, Issue 11
November 2008
Pages 773-775

Le problème

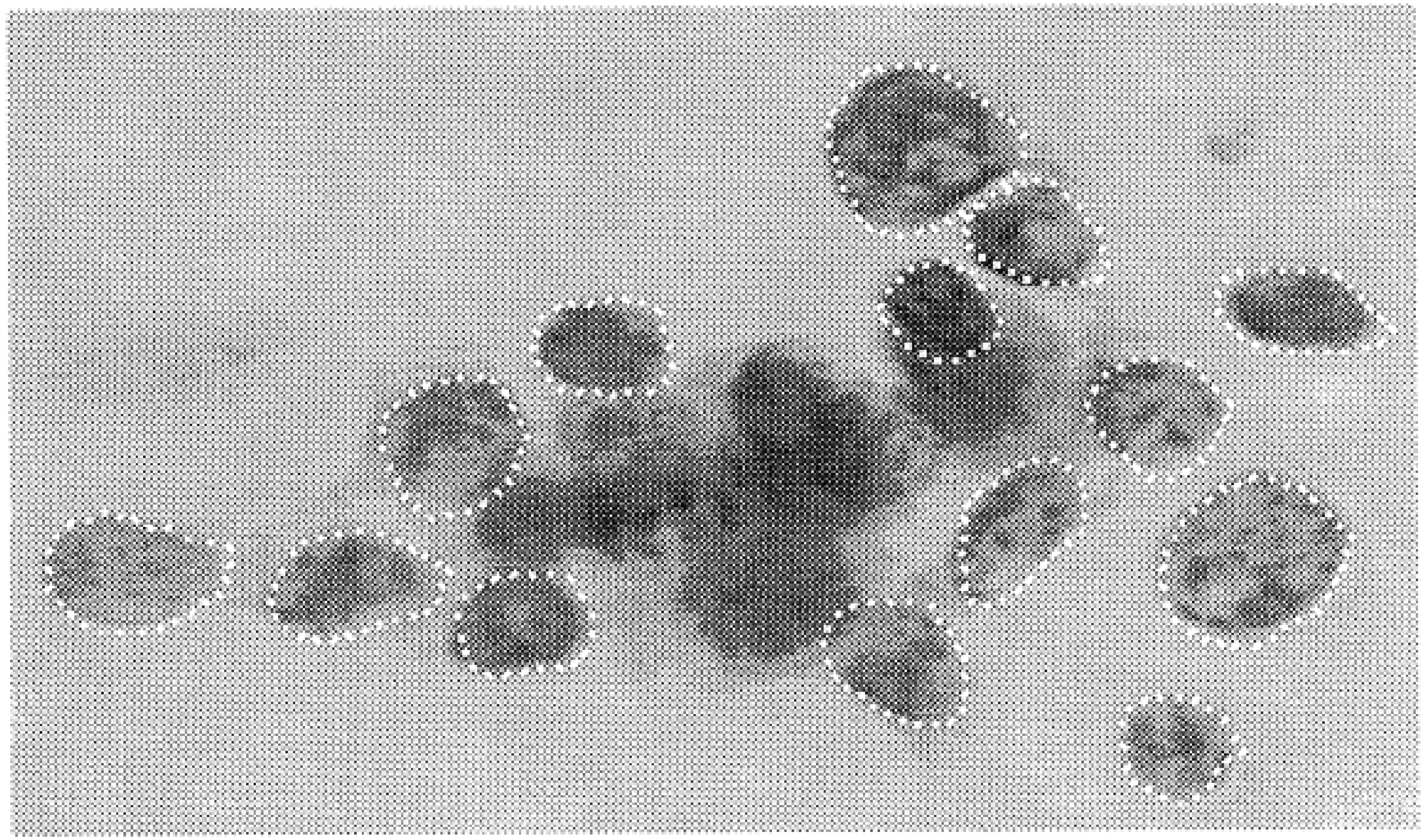


Figure 1: Initial Approximate Boundaries of Cell Nuclei
The user first draws a rough initial outline of some cell nucleus boundaries. Each outline serves as the initial position for a deformable spline which converges to an accurate boundary of the nucleus.

Le problème

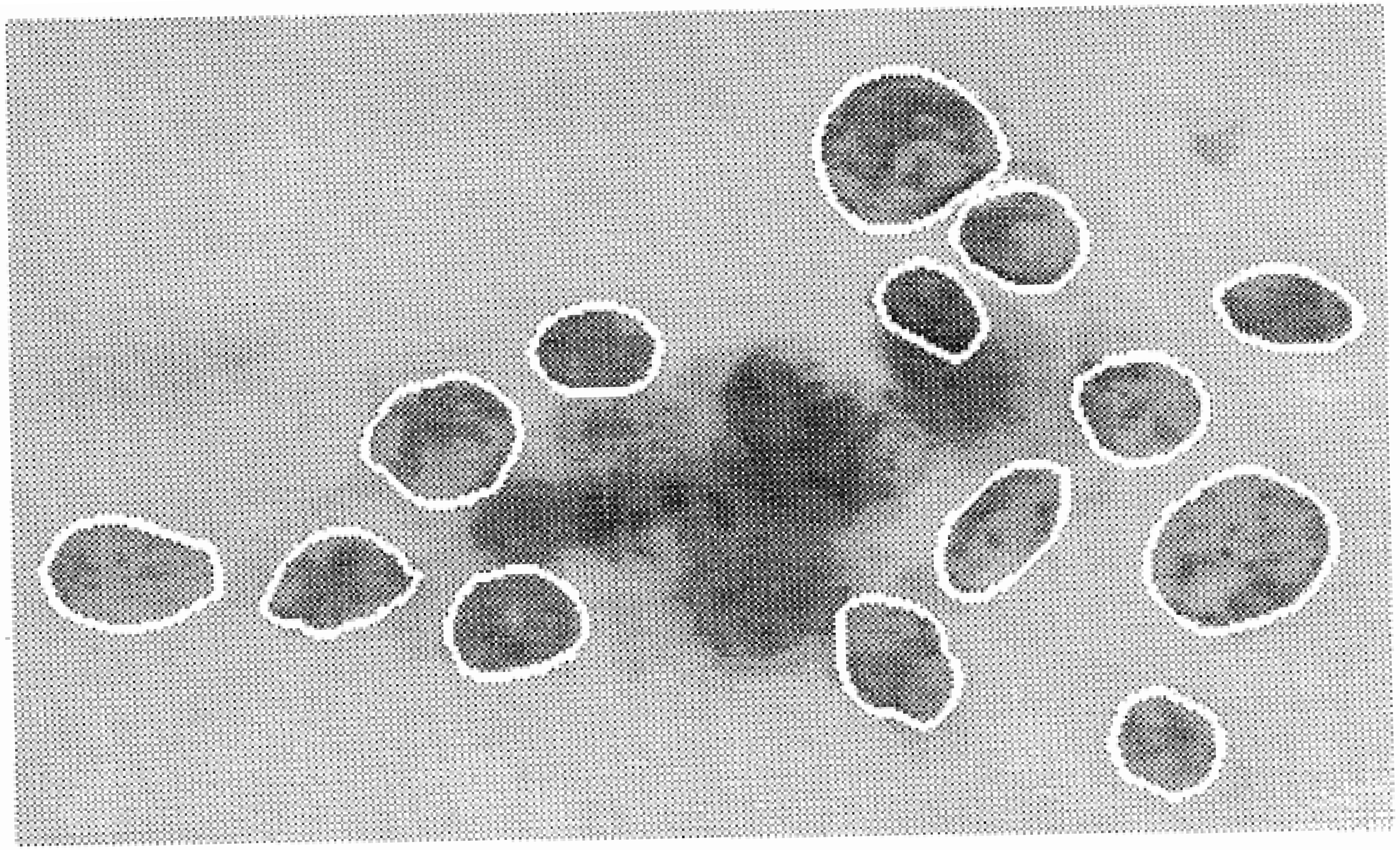


Figure 2: Snakes After Convergence to Cell Nucleus Boundaries
These contours are the final representation of the cell nuclei boundaries after the user is satisfied with the convergence of the snakes. This interactive process takes about two to five minutes.

Le dataset

Le dataset

Rayon

The radius of an individual nucleus is measured by averaging the length of the radial line segments defined by the centroid of the snake and the individual snake points.

Le dataset

Rayon

Texture

The texture of the cell nucleus is measured by finding the variance of the gray scale intensities in the component pixels.

Le dataset

Rayon

Texture

Périmètre

The total distance between the snake points constitutes the nuclear perimeter.

Le dataset

Rayon

Texture

Périmètre

Aire

Nuclear area is measured simply by counting the number of pixels on the interior of the snake and adding one-half of the pixels in the perimeter.

Le dataset

Rayon

Texture

Périmètre

Aire

Régularité

The smoothness of a nuclear contour is quantified by measuring the difference between the length of a radial line and the mean length of the lines surrounding it. This is similar to the curvature energy computation in the snakes. See Figure 3.

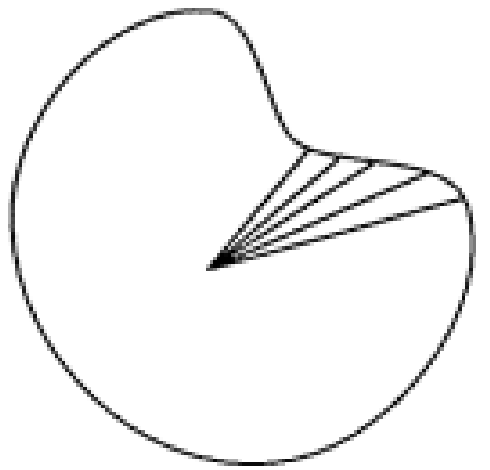


Figure 3: Radial Lines Used for Smoothness Computation

Le dataset

- Rayon
- Texture
- Périmètre
- Aire
- Régularité
- **Compacité**

Perimeter and area are combined[1] to give a measure of the compactness of the cell nuclei using the formula $perimeter^2/area$. This dimensionless number is minimized by a circular disk and increases with the irregularity of the boundary. However, this measure of shape also increases for elongated cell nuclei, which do not necessarily indicate an increased likelihood of malignancy. The feature is also biased upward for small cells because of the decreased accuracy imposed by digitization of the sample. We compensate for the fact that no single shape measurement seems to capture the idea of "irregular" by employing several different shape features.

Le dataset

- Rayon
- Texture
- Périmètre
- Aire
- Régularité
- Compacité
- **Concavité**

In a further attempt to capture shape information we measure the number and severity of concavities or indentations in a cell nucleus. We draw chords between non-adjacent snake points and measure the extent to which the actual boundary of the nucleus lies on the inside of each chord (see Figure 4). This parameter is greatly affected by the length of these chords, as smaller chords better capture small concavities. We have chosen to emphasize small indentations, as larger shape irregularities are captured by other features.

Figure 4: Chords Used to Compute Concavity

Le dataset

- Rayon
- Texture
- Périmètre
- Aire
- Régularité
- Compacité
- Concavité
- **Points concaves**

This feature is similar to Concavity but measures only the number, rather than the magnitude, of contour concavities.

Le dataset

- Rayon
- Texture
- Périmètre
- Aire
- Régularité
- Compacité
- Concavité
- Points concaves
- **Symétrie**

In order to measure symmetry, the major axis, or longest chord through the center, is found. We then measure the length difference between lines perpendicular to the major axis to the cell boundary in both directions. See Figure 5. Special care is taken to account for cases where the major axis cuts the cell boundary because of a concavity.

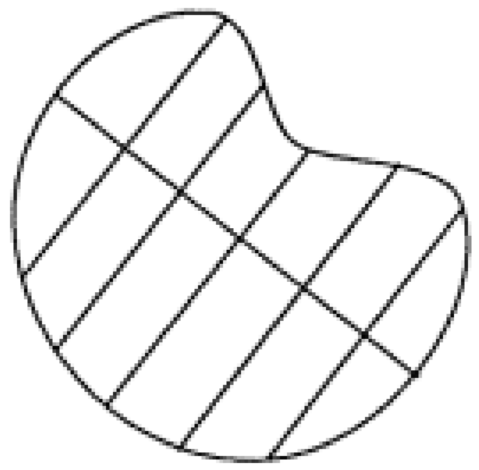


Figure 5: Segments Used in Symmetry Computation

Le dataset

- Rayon
- Texture
- Périmètre
- Aire
- Régularité
- Compacité
- Concavité
- Points concaves
- Symétrie
- **Dimension fractale**

The fractal dimension of a cell is approximated using the "coastline approximation" described by Mandelbrot.[9] The perimeter of the nucleus is measured using increasingly larger 'rulers'. As the ruler size increases, decreasing the precision of the measurement, the observed perimeter decreases. See Figure 6. Plotting these to values on a log scale and measuring the downward slope gives (the negative of) an approximation to the fractal dimension. As with all the shape features, a higher value corresponds to a less regular contour and thus to a higher probability of malignancy.

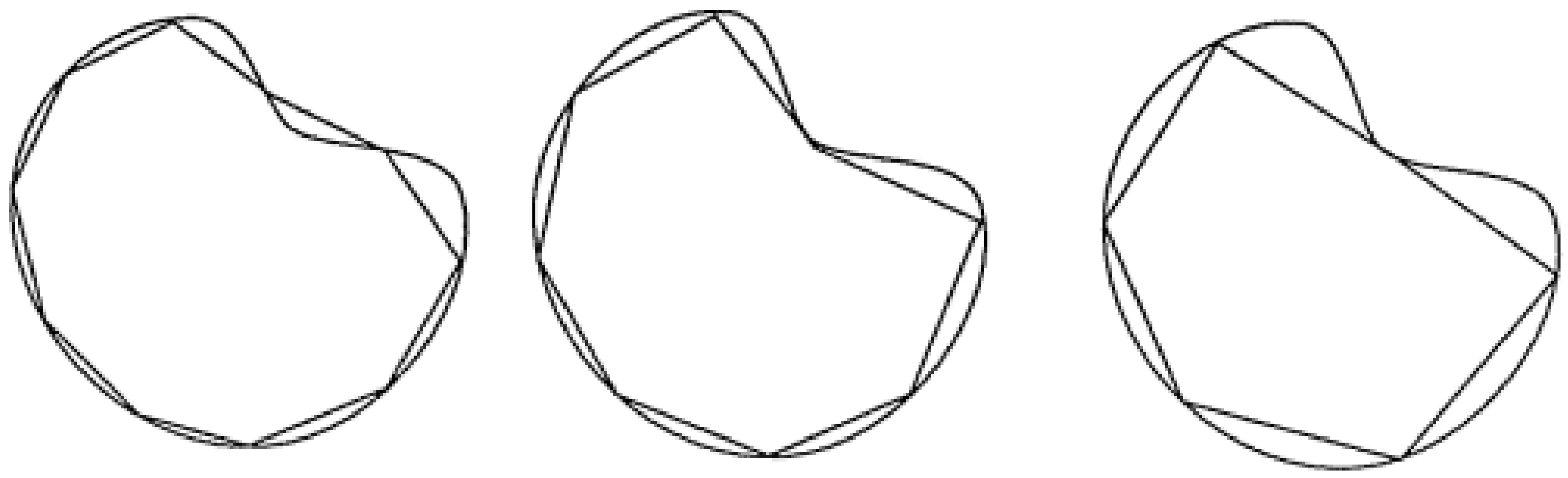
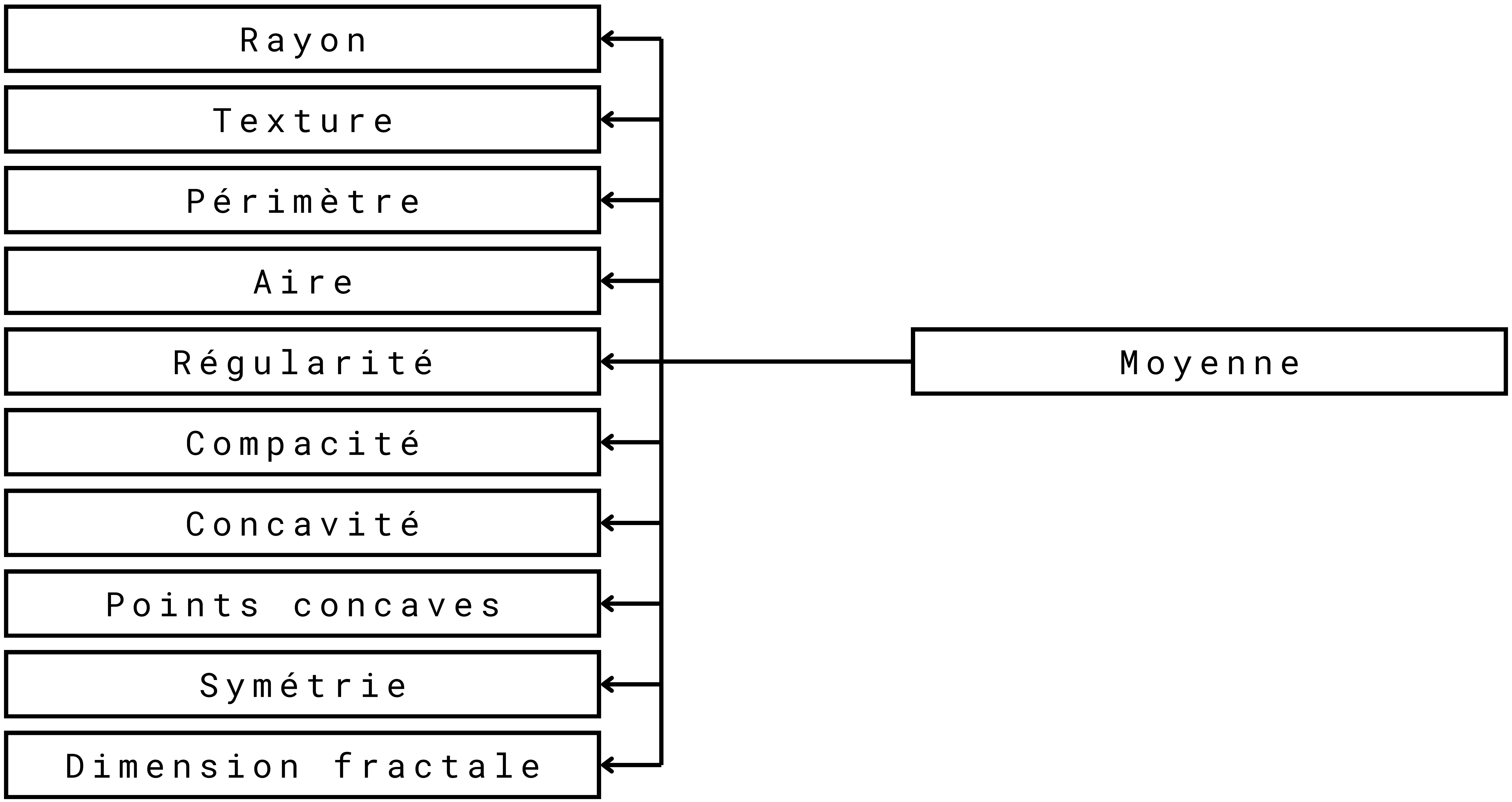


Figure 6: Sequence of Measurements for Computing Fractal Dimension

Le dataset

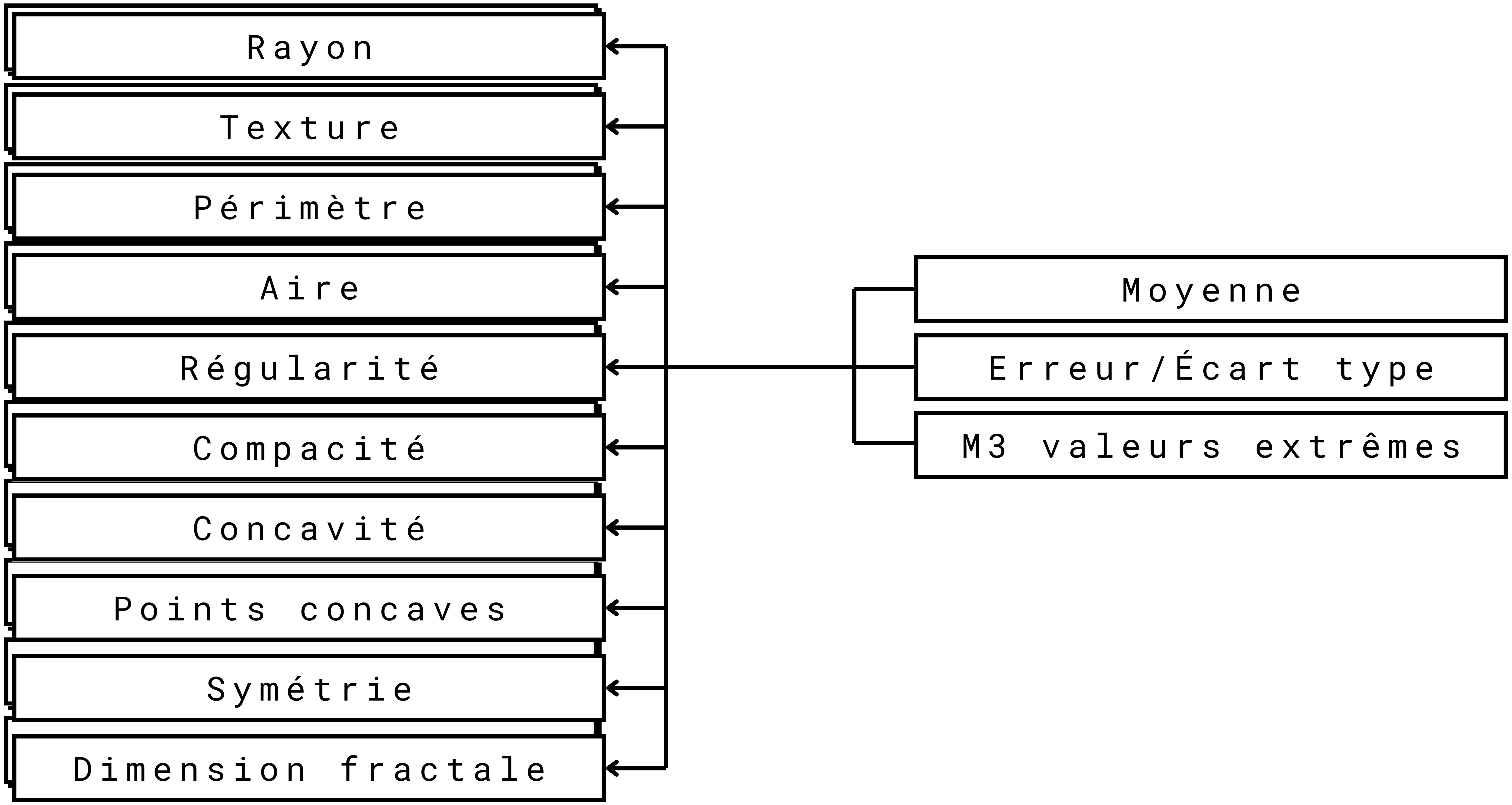
- Rayon
- Texture
- Périmètre
- Aire
- Régularité
- Compacité
- Concavité
- Points concaves
- Symétrie
- Dimension fractale

Le dataset

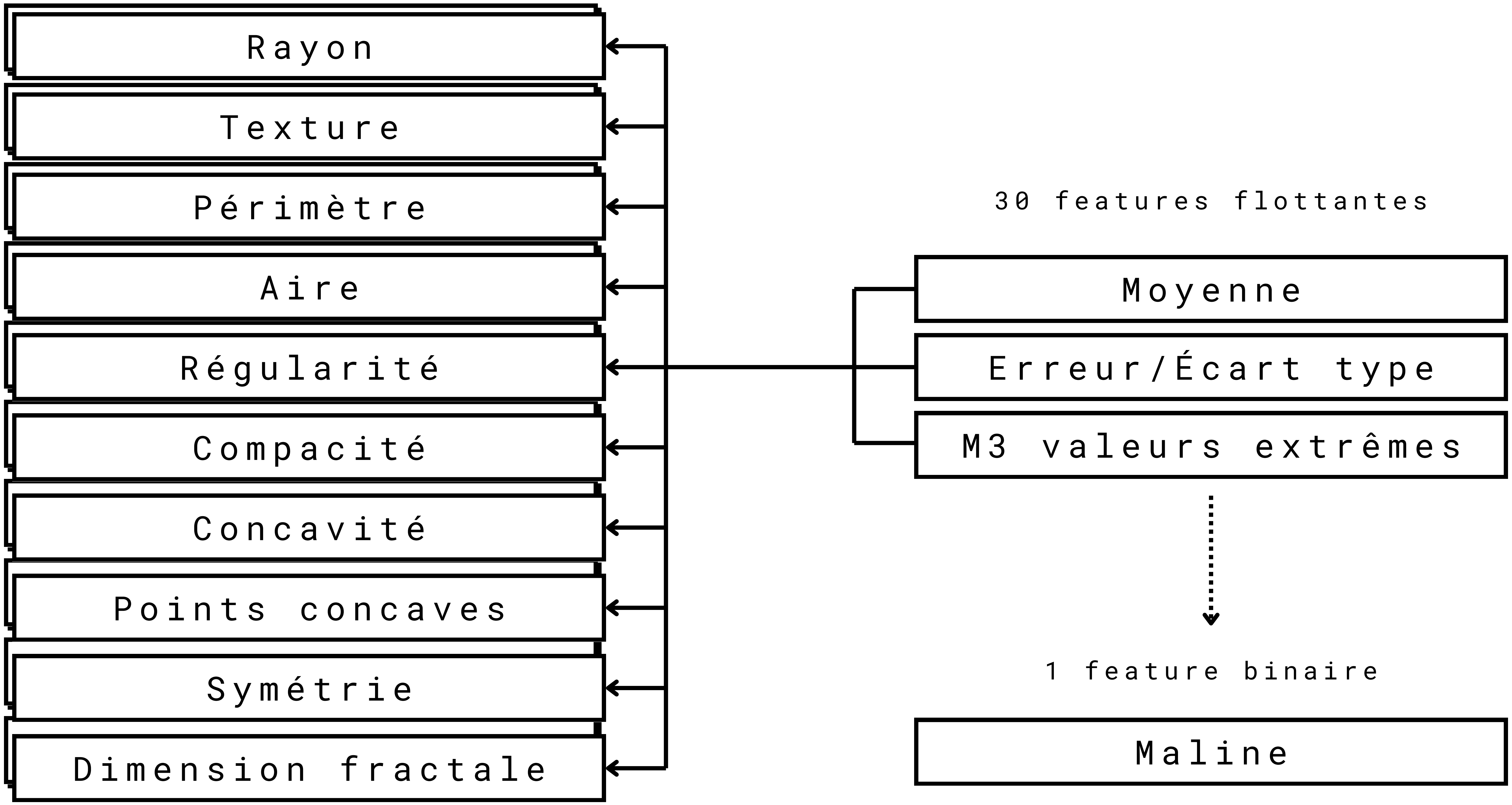


Rayon
Texture
Périmètre
Aire
Régularité
Compacité
Concavité
Points concaves
Symétrie
Dimension fractale
Moyenne

Le dataset


Rayon
Texture
Périmètre
Aire
Régularité
Compacité
Concavité
Points concaves
Symétrie
Dimension fractale
Moyenne
Erreur/Écart type
M3 valeurs extrêmes

Le dataset


Rayon
Texture
Périmètre
Aire
Régularité
Compacité
Concavité
Points concaves
Symétrie
Dimension fractale
30 features flottantes
Moyenne
Erreur/Écart type
M3 valeurs extrêmes
1 feature binaire
Maline

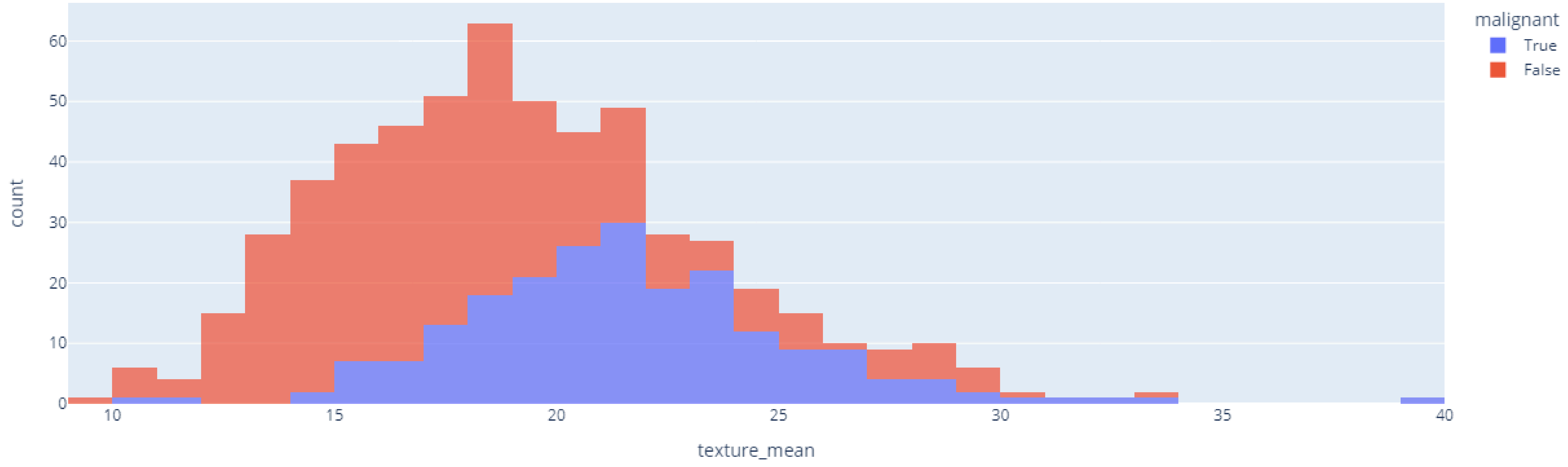
> > > > > > > >

Let's work !

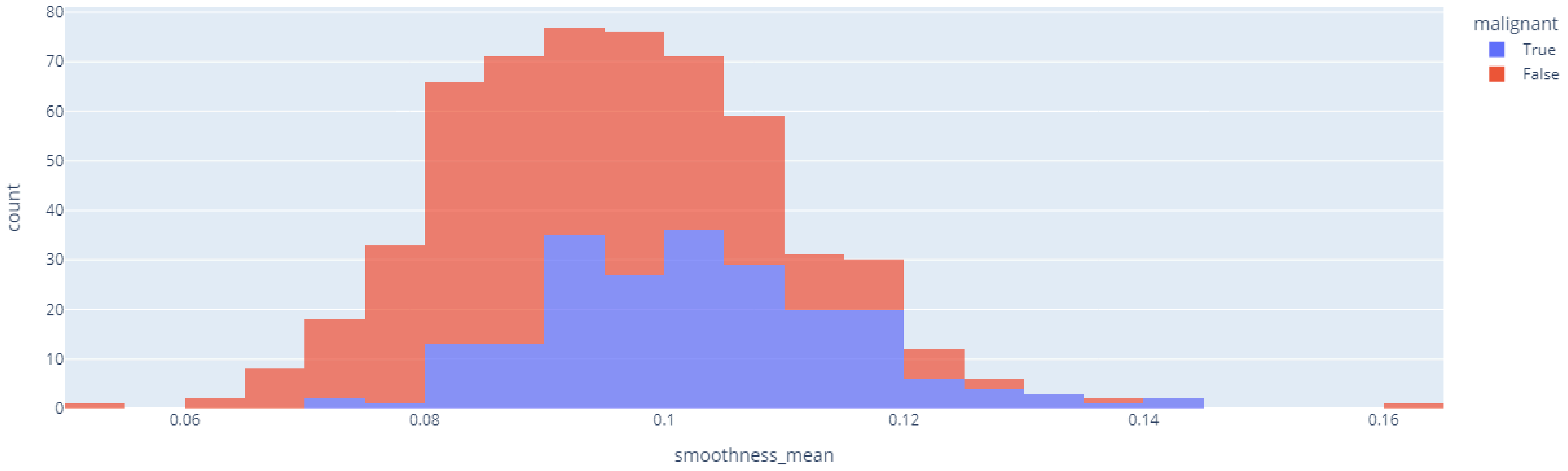
Au boulot !

Au travail !

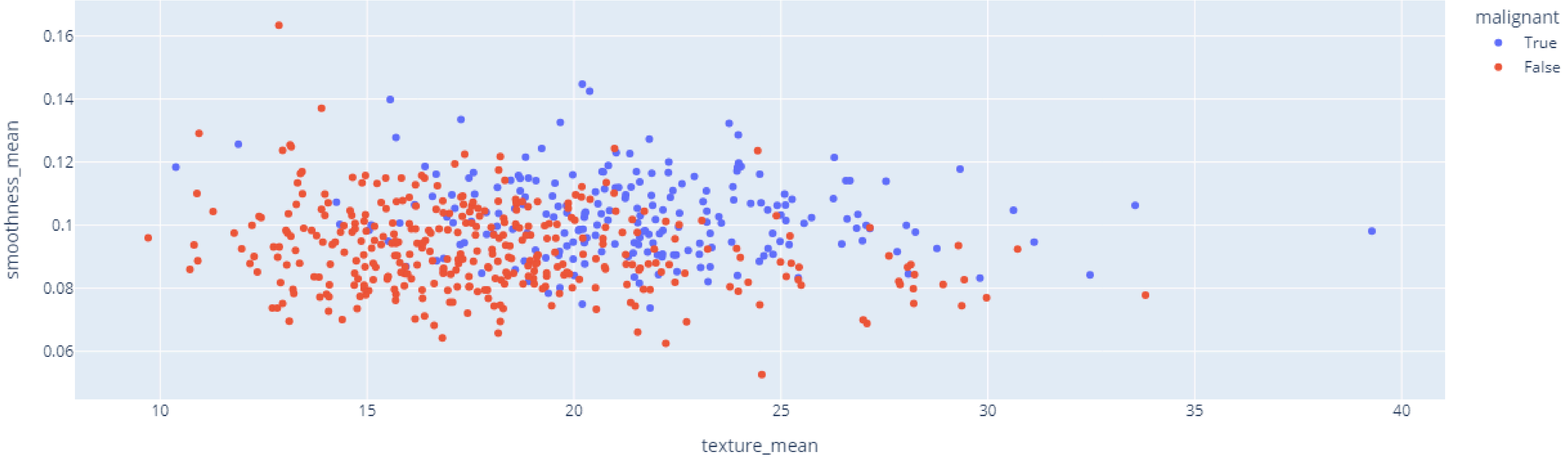
Analyse des données



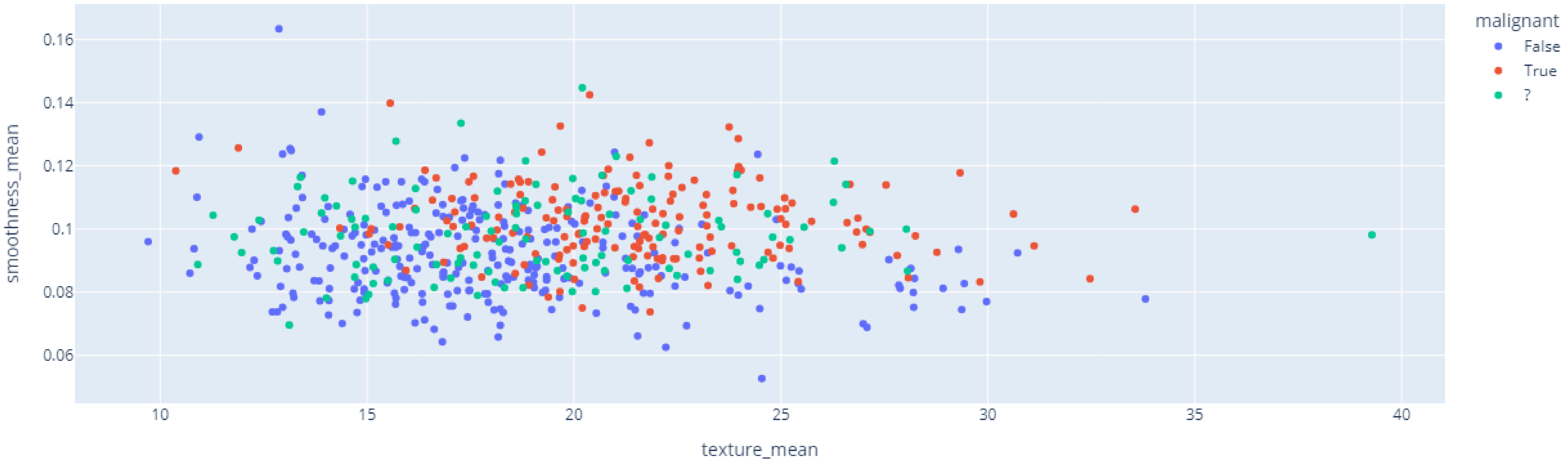
Analyse des données



Analyse des données

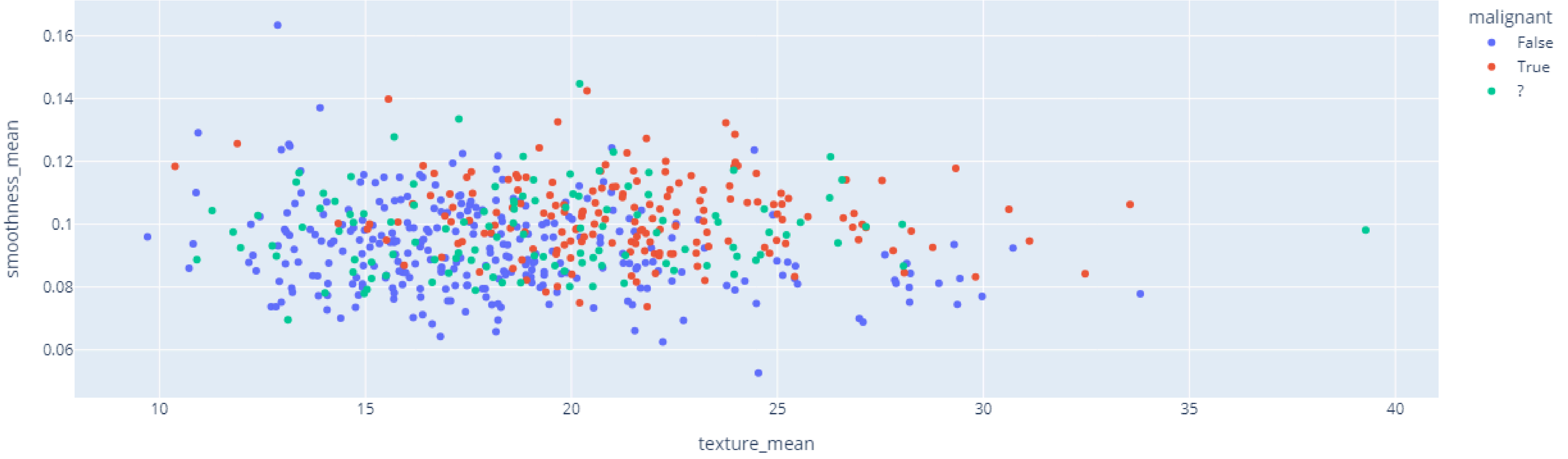


Analyse des données



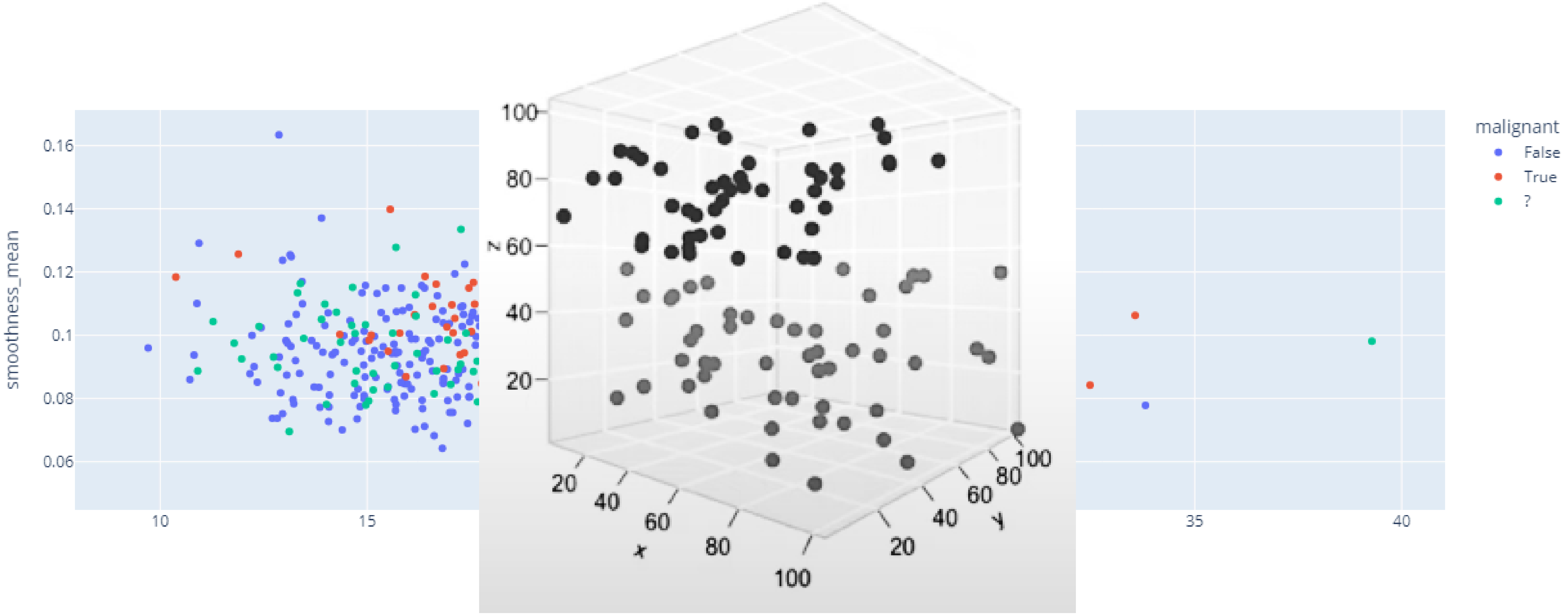
kNN

kNN



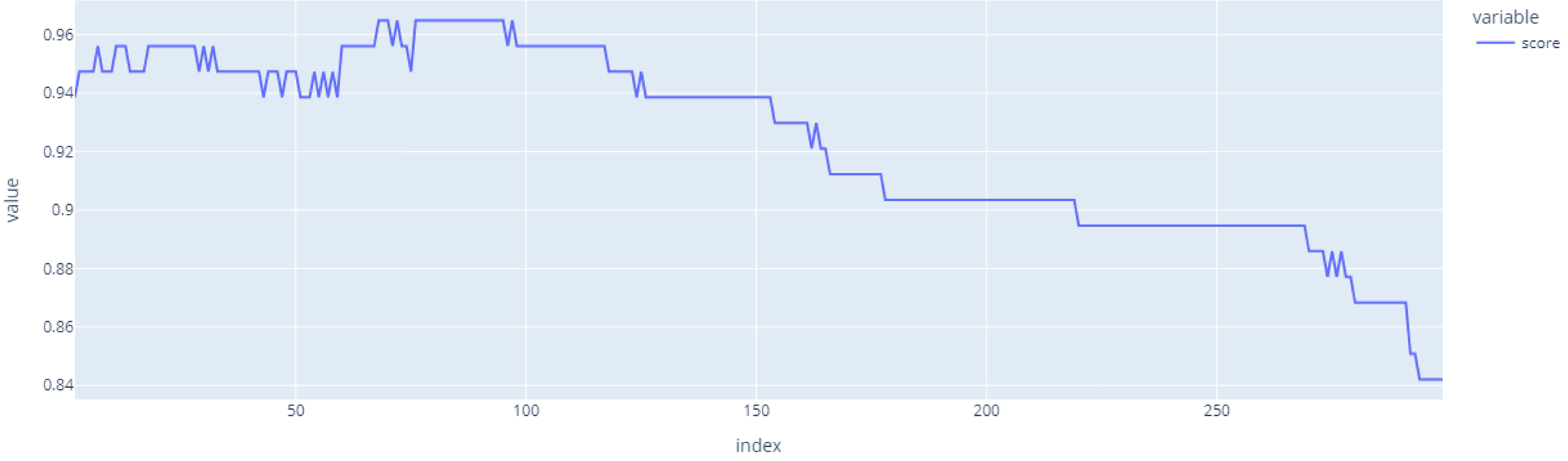
malignant
False
True
?
smoothness_mean
texture_mean
0.16
0.14
0.12
0.1
0.08
0.06
10
15
20
25
30
35
40

kNN

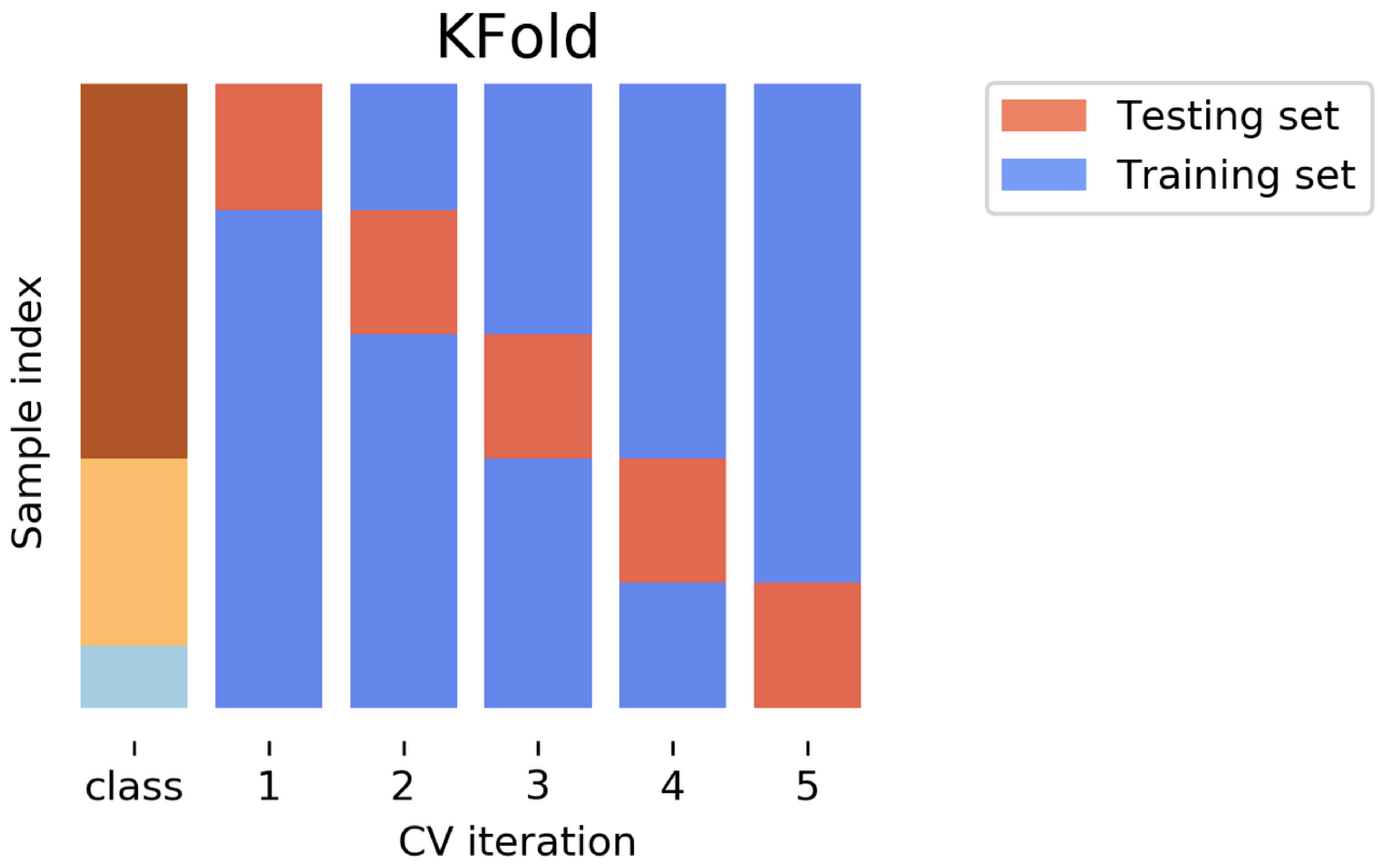


smoothness_mean
0.16
0.14
0.12
0.1
0.08
0.06
10
15
35
40
malignant
False
True
?
100
80
60
40
20
z
x
y

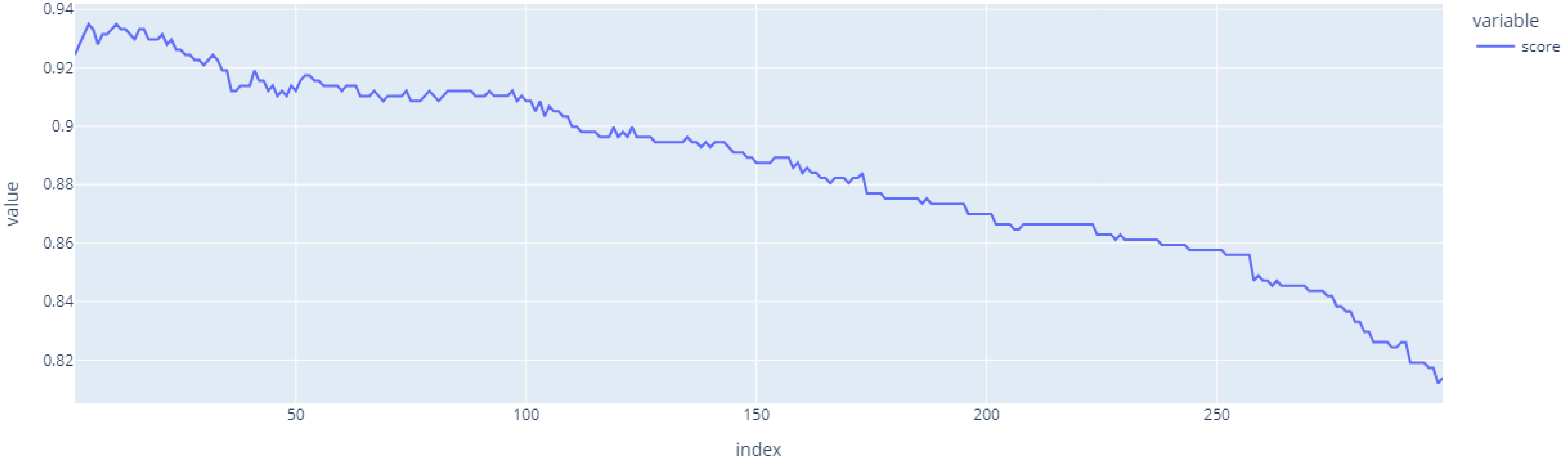
kNN > Hyper Paramètres



Validation croisée



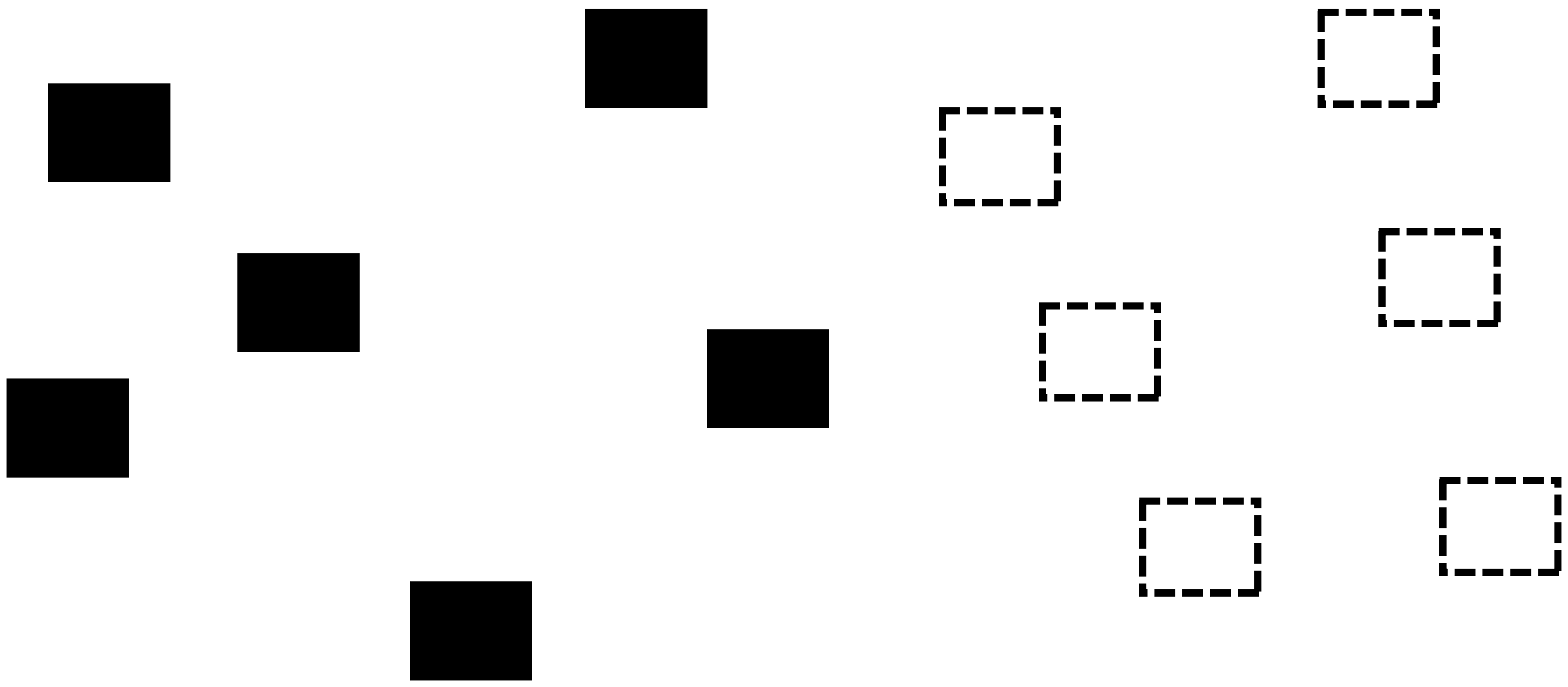
kNN > Validation croisée



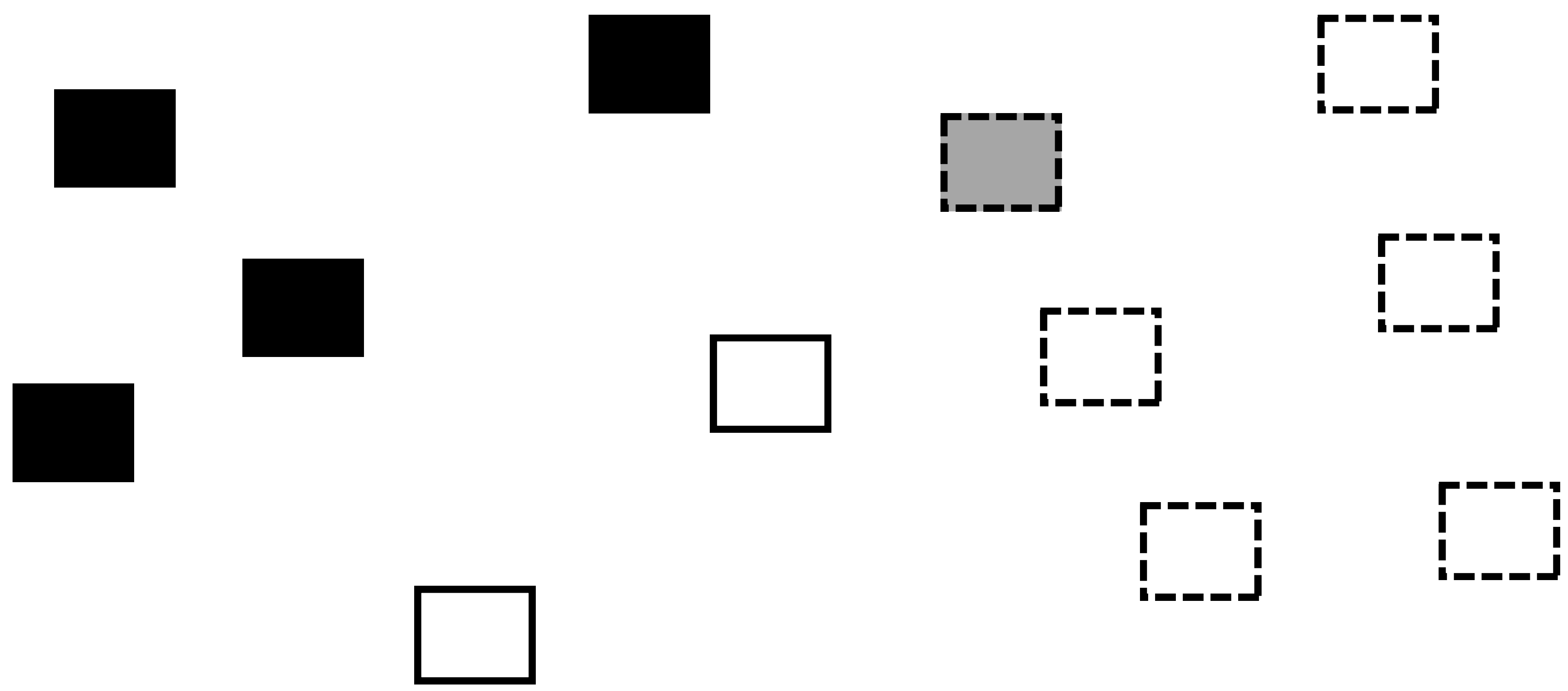
Précision, Rappel, F1

Exactitude / Accuracy

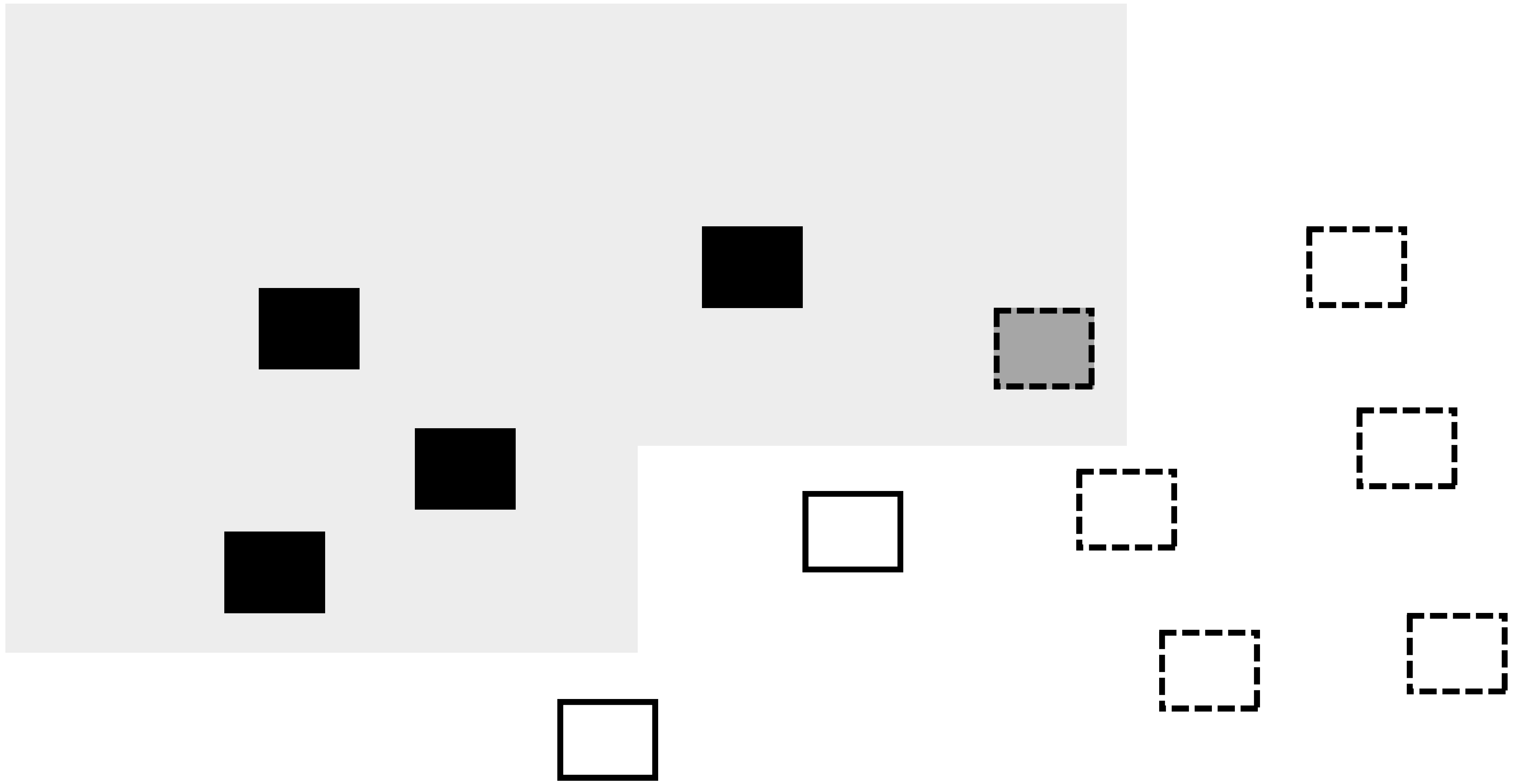
Précision, Rappel, F1



Précision, Rappel, F1



Précision, Rappel, F1



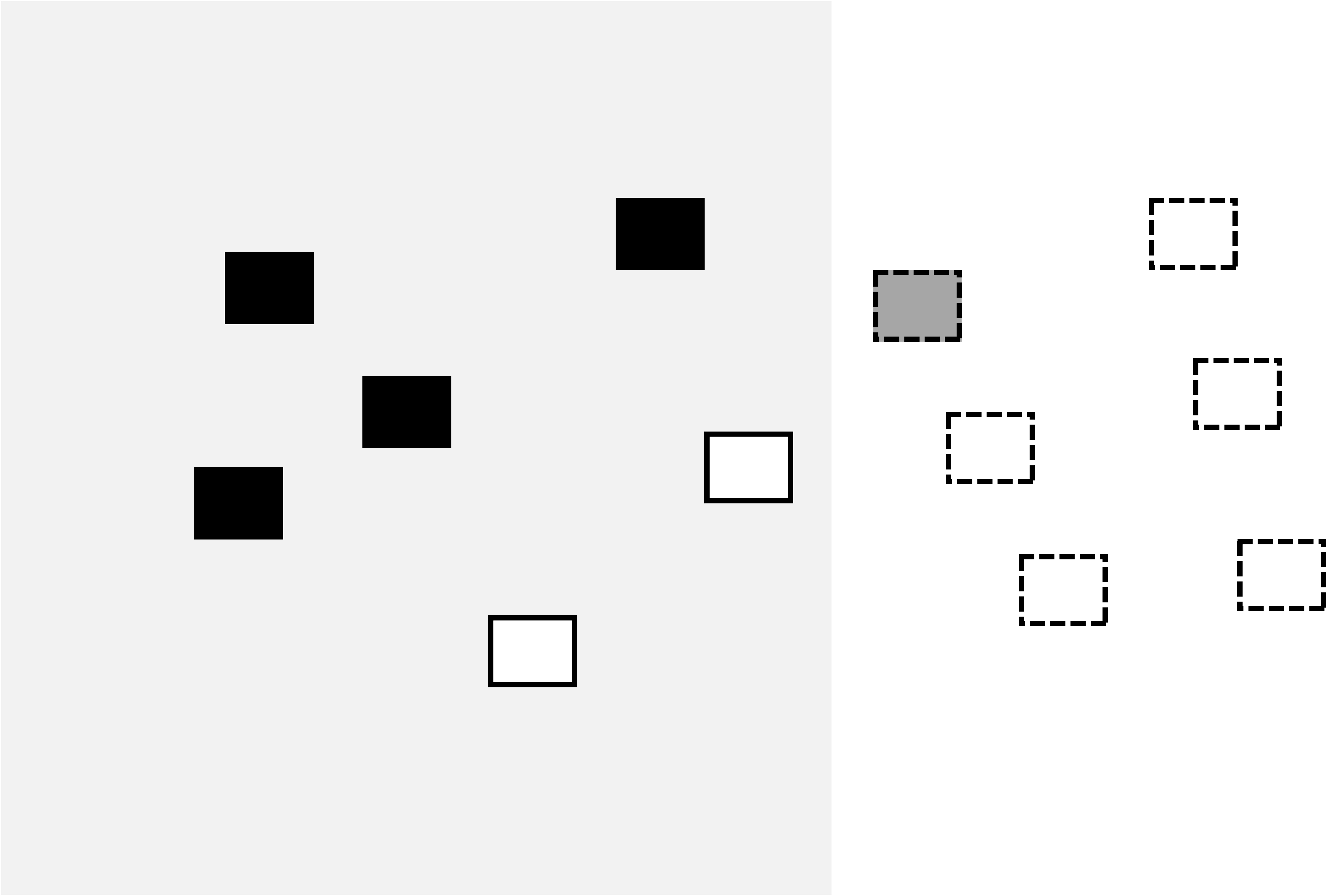
Précision :

VP/(VP + FP)

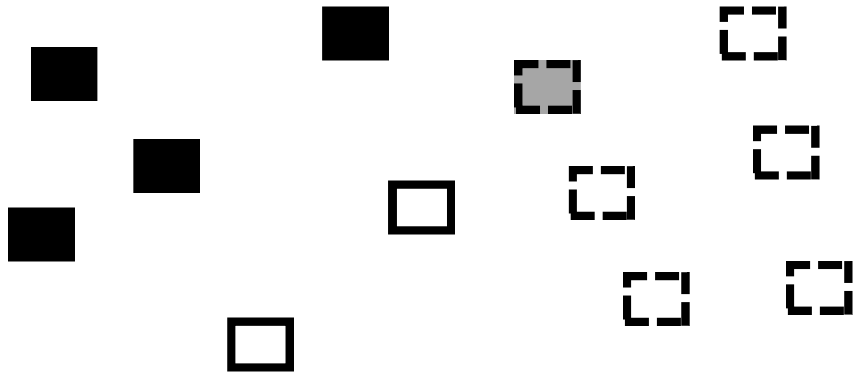
Précision, Rappel, F1

Rappel :

VP/(VP + FN)



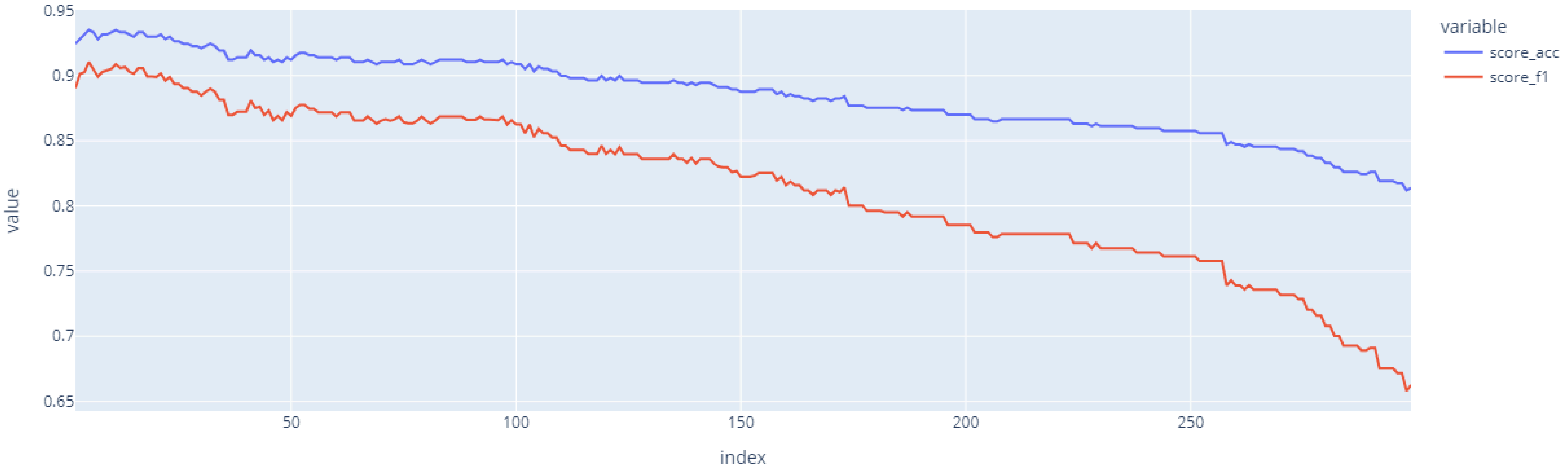
Précision, Rappel, F1



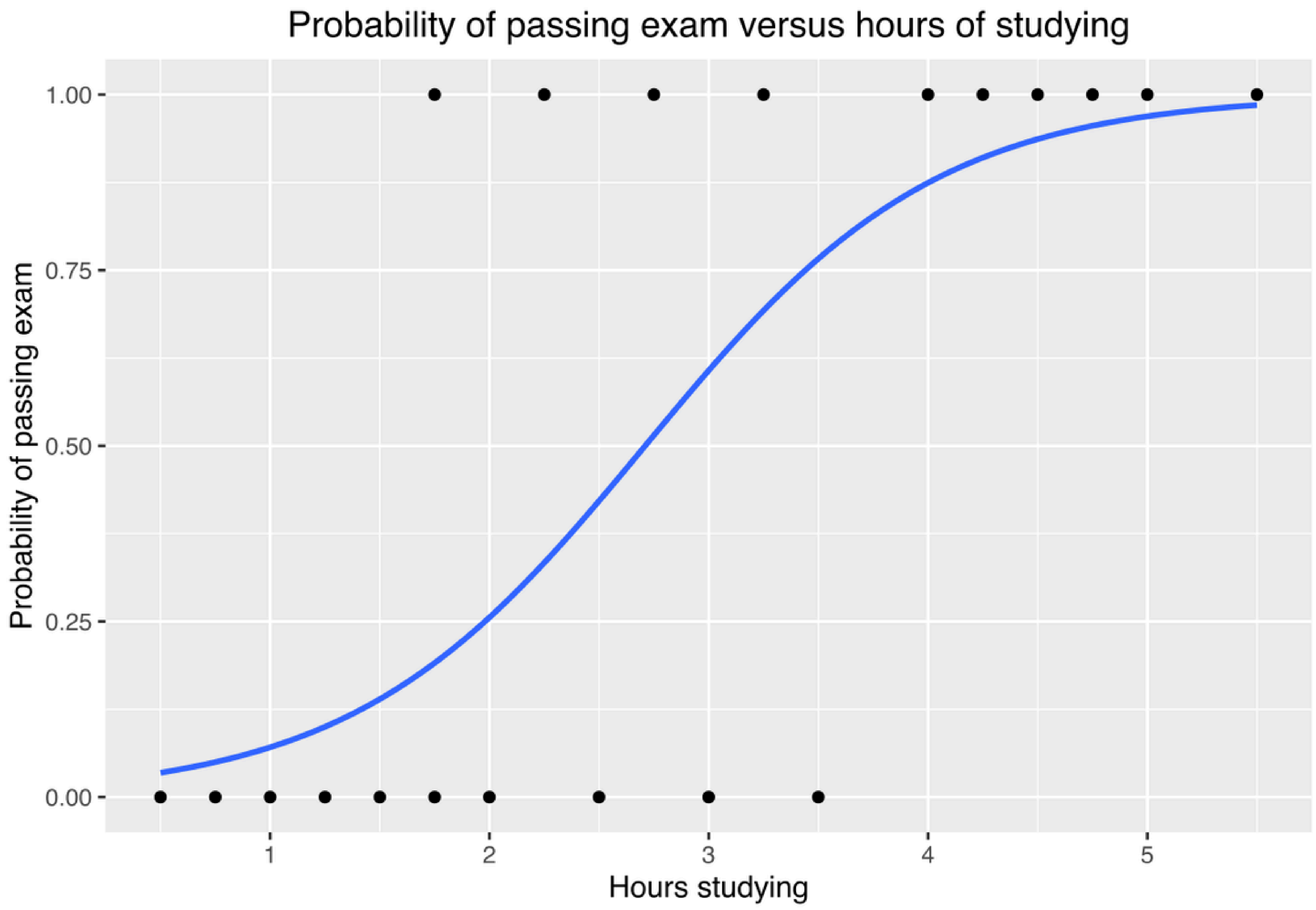
$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

$$= \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

kNN > F1 cross-validation



Regression logistique



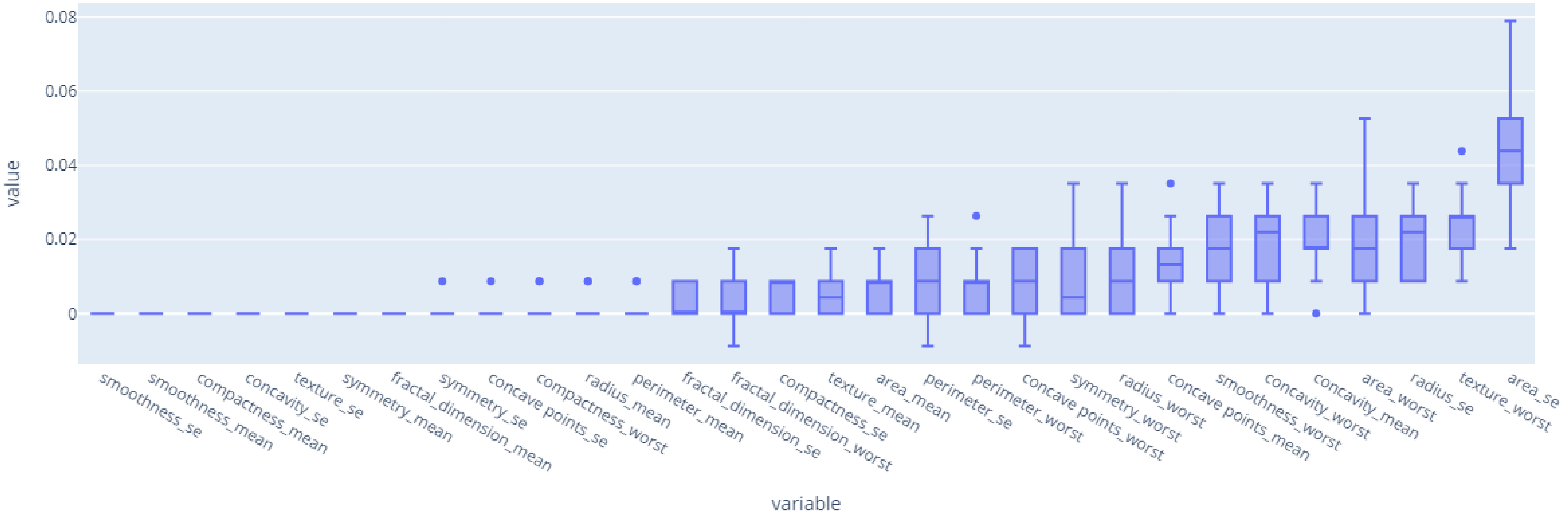
Standardisation, normalisation

$$z = \frac{x - \mu}{\sigma}$$

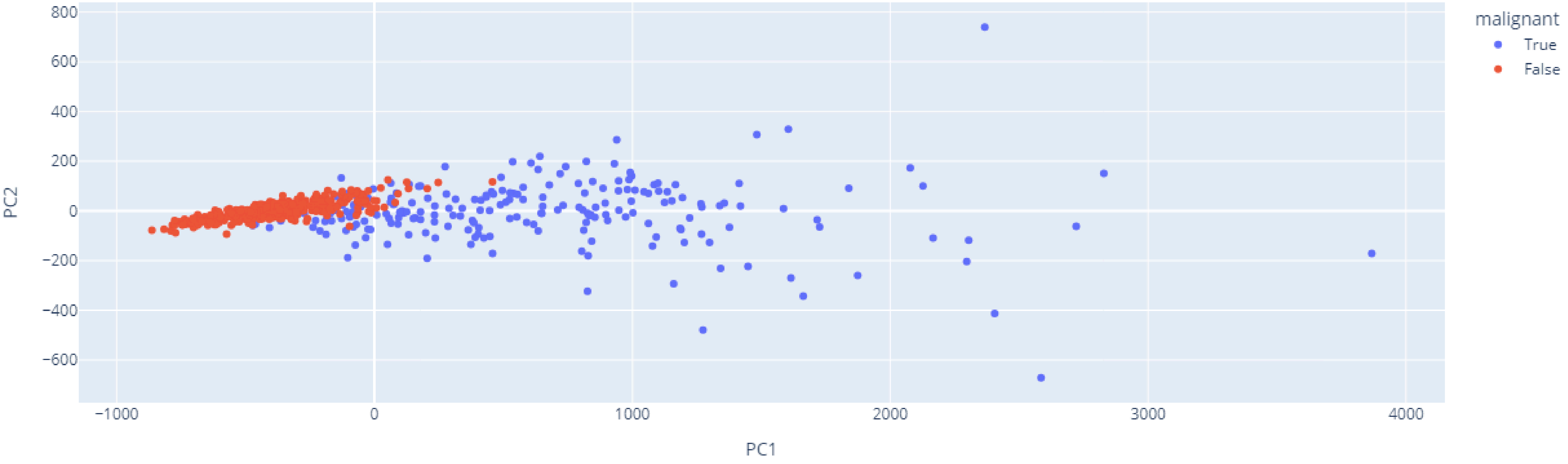
Standardisation, normalisation

$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$

Interprétatibilité



ACP



800
600
400
200
0
−200
−400
−600
PC2
−1000
0
1000
2000
3000
4000
PC1
malignant
True
False

[Et voilà...]

Fin de l’atelier :)

SIAHAAN--GENSOLLEN Rémy, 2025