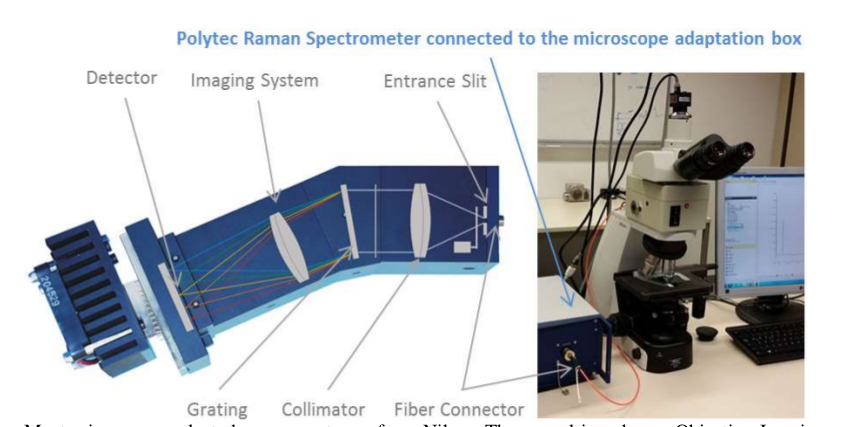
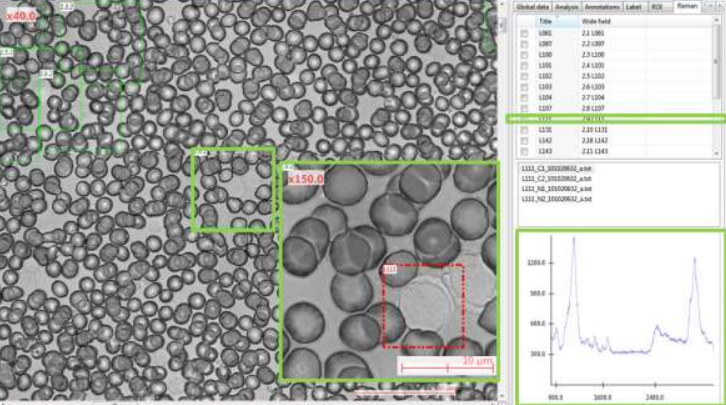
**Scénario**:

Notre client, un hôpital, désire optimiser le processus de diagnostic de la Leucémie Lymphoïde Chronique (LLC), maladie cancéreuse qui se traduit par une prolifération monoclonale de **lymphocytes B** qui vont envahir progressivement le sang. Pour ce faire cet hôpital dispose de données issues de la spectrométrie Raman.



Le spectromètre Raman pointe sur trois régions aléatoires d’une cellule et génère les spectres associés à chacun de ces points :



Les spectres passent ensuite une série de contrôles de qualité :

« The spectral data are pre-processed for correction of spectral interferences [...]. Such pre-processed spectra are submitted to quality tests such as signal to noise ratio, spikes. Spectra fulfilling these quality tests are kept for further numerical/statistical analysis »

Uniquement les spectres passant les contrôles sont conservés, conduisant à un nombre différent de cellules par patient et de spectres par cellule.

L’objectif de ce cas technique est de **développer un algorithme permettant de diagnostiquer la LLC en prédisant la proportion de cellules B** pour un patient donné. Celui-ci sera utilisé par l’hôpital pour assister le travail des laborantins.

**Données :**

* **raw\_data.csv**: une table en 1 ligne par spectre, répertoriant les informations suivantes :
  + **patient\_name** : Identifiant unique du patient (identifiant du patient)
  + **cell\_name**: Identifiant unique de la cellule (identifiant de la cellule)
  + **cell\_type**: type de la cellule (**lymphocyte B** ou **lymphocyte TNK**)
  + **patient\_state**: Booléen : état du patient (malade ou sain)
  + **spectre** : numéro du spectre pour une cellule donnée (**identifiant arbitraire du spectre**)
  + **lambda\_1 à lambda\_999**: amplitude mesurée pour chaque longueur d’onde.

**N.B.** patient\_name et cell\_name sont des identifiants uniques et arbitraires des patients et cellules.

**Travaux à réaliser**

Dans ce cadre, nous vous demandons de détailler les éléments suivants :

1. Analyse descriptive des données et les conclusions que vous en tirez
2. Méthodologie de la structuration des données effectuée
3. Détail des algorithmes et stratégies analytiques envisagés et retenues
4. Présentation type ppt des résultats obtenus et des recommandations associées. La qualité attendue est identique à celle que vous pourriez livrer au client lui-même

**Remarques** :

Il est attendu que le code (R ou python) et la présentation (ppt) soient **envoyés 24h au préalable**.