# Classification of biological cells to cancer and non-cancer ones by their surface-enhanced Raman scattering

Anastasiia Merdalimova

February 2023

# Contents

- Introduction to the subject
- Problem statement
- Data research
- Applying machine learning techniques
  - Dimensionality reduction
  - Classification
- Task modification

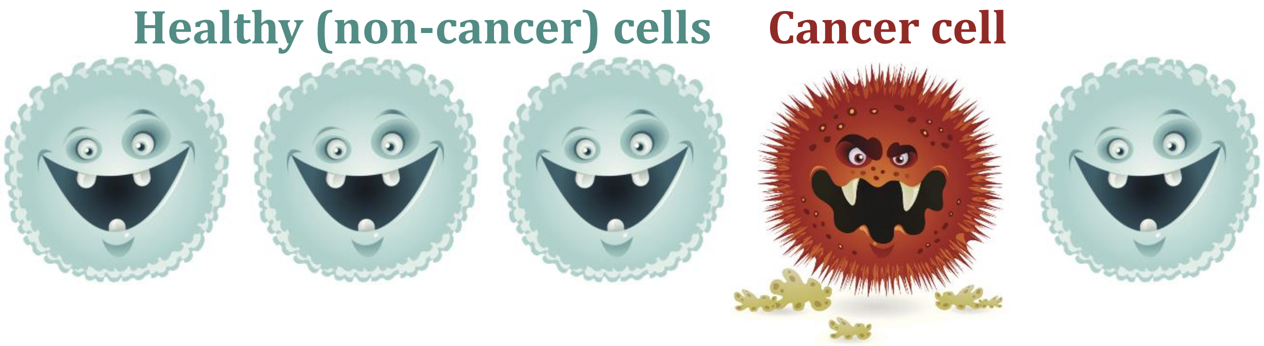
# Contents

- **Introduction to the subject**
- Problem statement
- Data research
- Applying machine learning techniques
  - Dimensionality reduction
  - Classification
- Task modification

# Subject

- **What?** Diagnosing cancer cells at the early stage

A problem of early-stage cancer detection
is currently one of hot topics in diagnostics

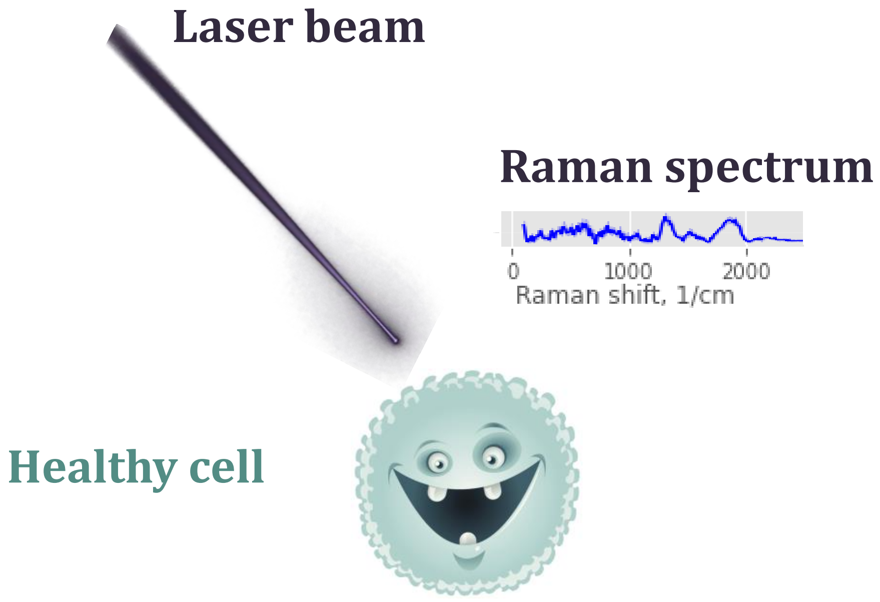


# Subject

## 1 Raman spectroscopy

BUT Signal intensity is weak :(

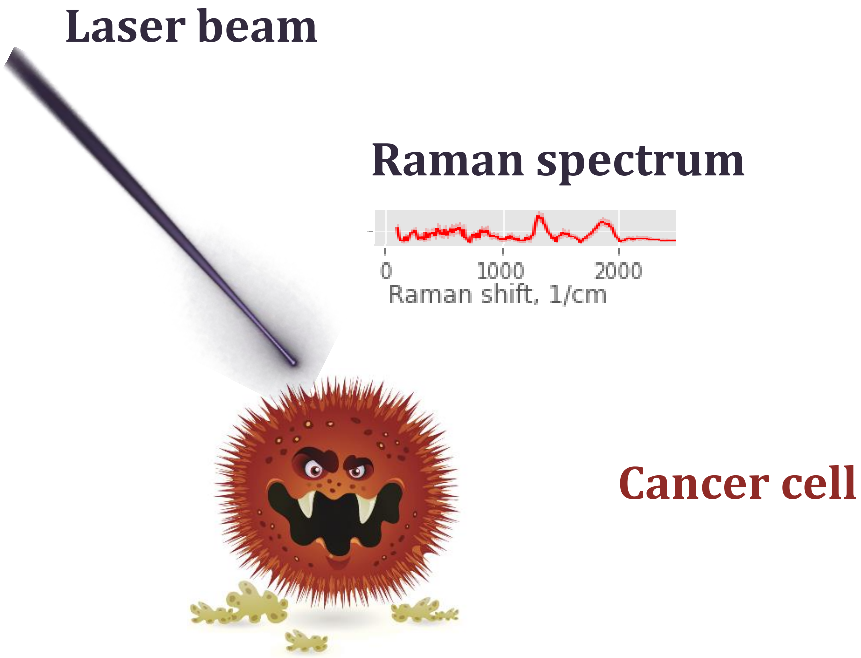
- **How?**



Laser beam

Raman spectrum

Healthy cell



Laser beam

Raman spectrum

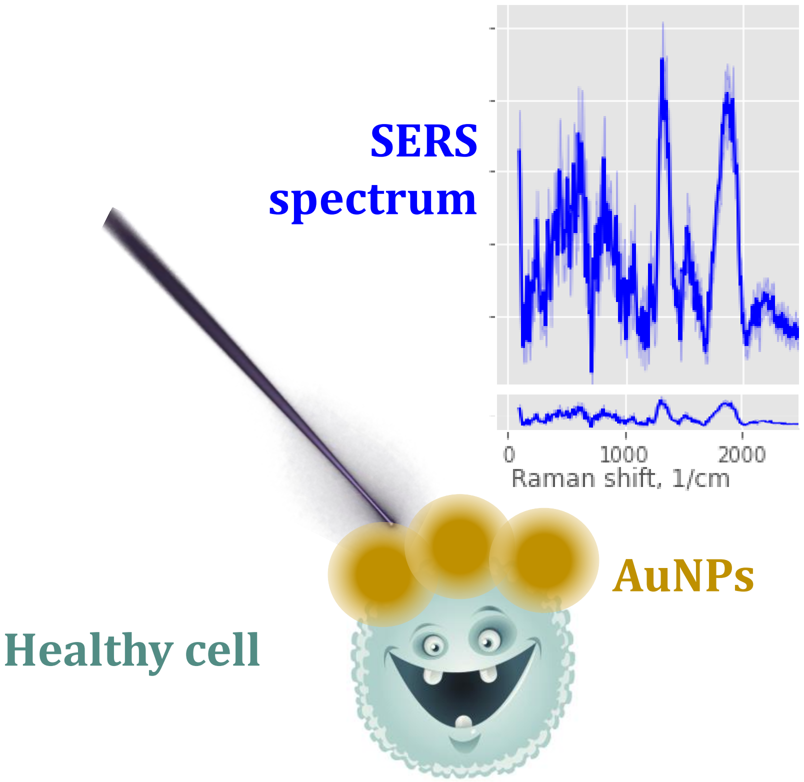
Cancer cell

Raman spectroscopy is one of methods of noninvasive analysis of substance contents, inter alia cell analysis. However, Raman signal is typically too low…

# Subject

- **How?** 1 **Raman spectroscopy**
  2 **+ enhancing gold particles (AuNPs)** BUT

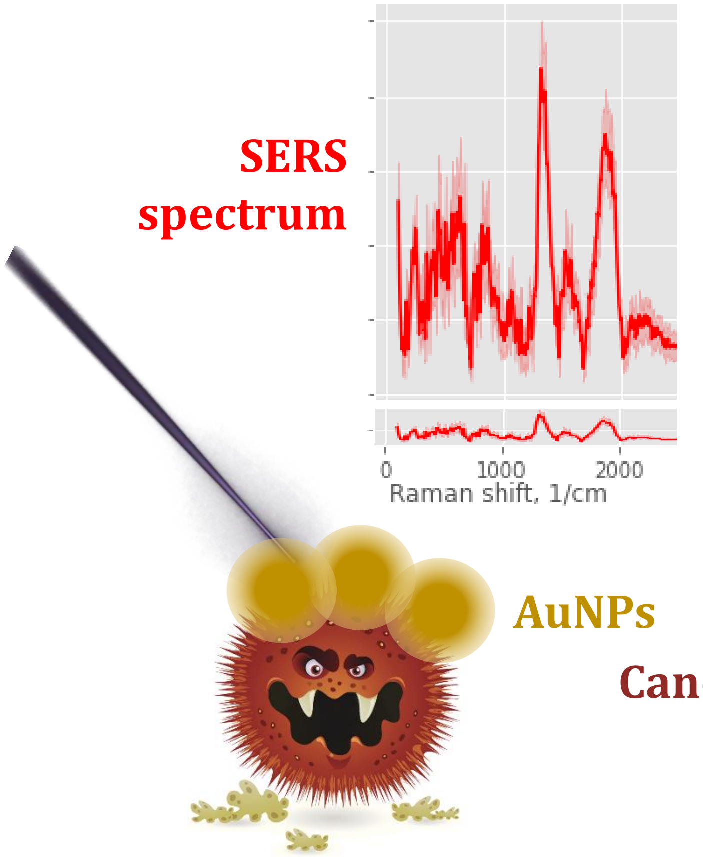
Provide enhancement only in their vicinity :(



**SERS spectrum**

**AuNPs**

**Healthy cell**



**SERS spectrum**

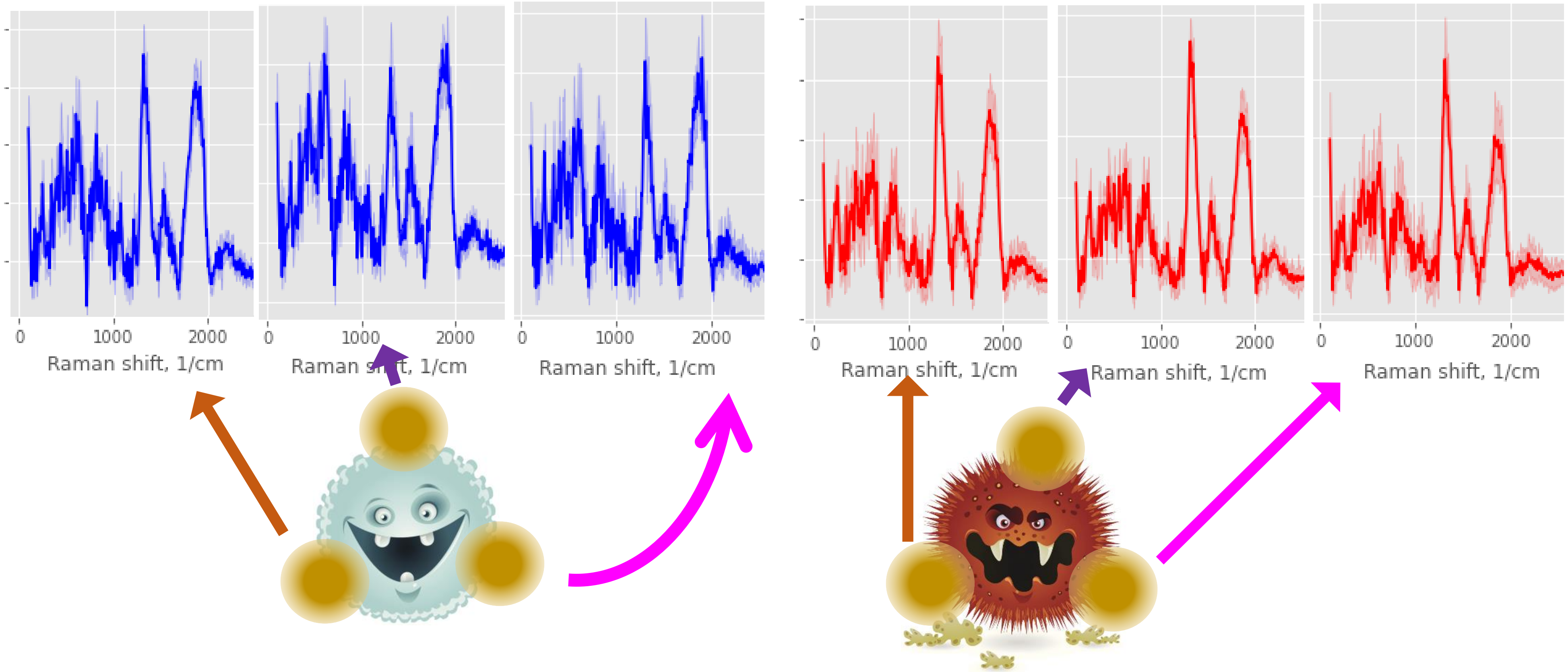
**AuNPs**

**Cancer cell**

… Therefore, special enhancing gold or silver nanoparticles may be used to enhance the signal; such approach is called SERS, Surface-Enhanced Raman Spectroscopy.

# Subject

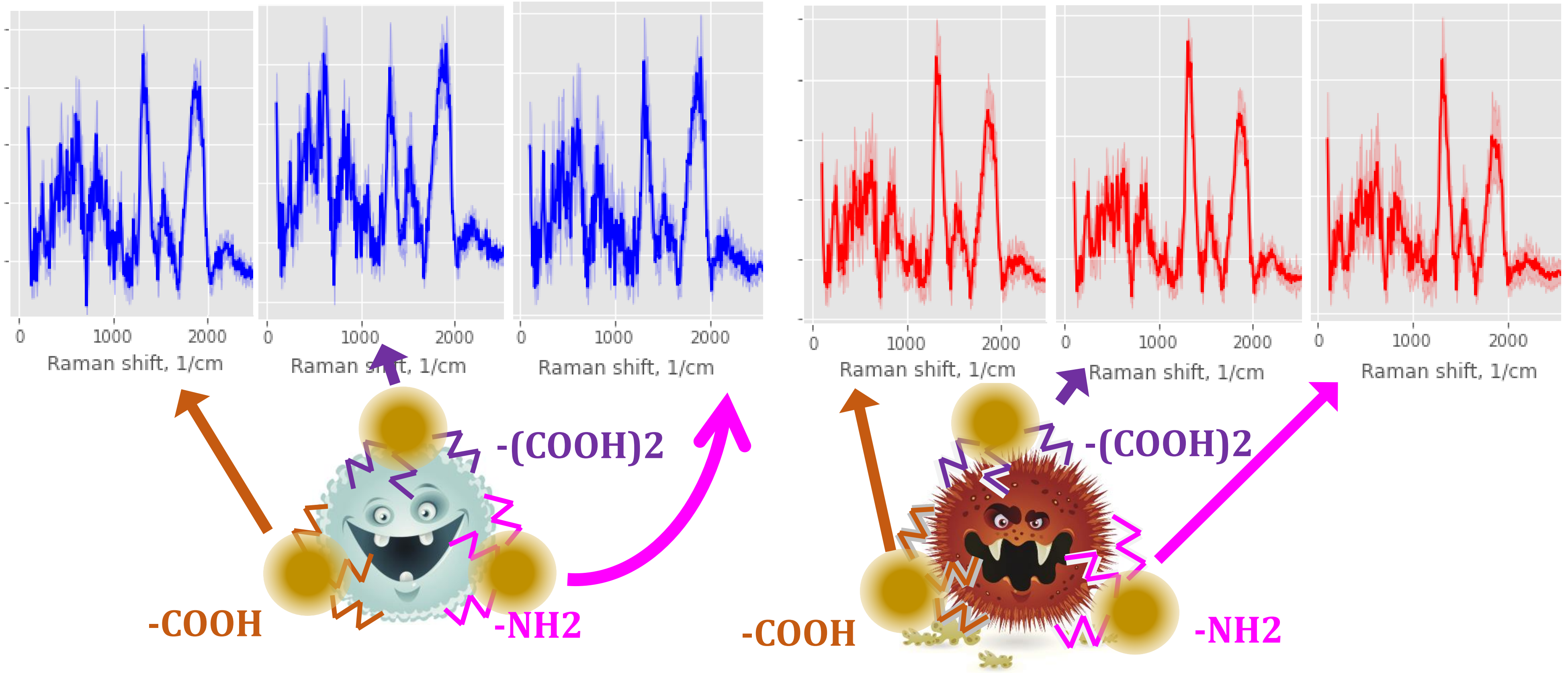
- **How?**
  1. **Raman spectroscopy**
  2. **+ enhancing gold particles (AuNPs)**
  3. **of three different types!**



The 3 types of AuNPs are functionalized with different chemical groups,

# Subject

- **How?**
  1. **Raman spectroscopy**
  2. **+ enhancing gold particles (AuNPs)**
  3. **of three different types!**



The 3 types of AuNPs are functionalized with different chemical groups, therefore bind to differect molecules and therefore provide different spectra from their vicinities.

# Contents

- Introduction to the subject
- **Problem statement**
- Data research
- Applying machine learning techniques
    - Dimensionality reduction
    - Classification
- Task modification

# Task

- To distinguish between healthy and cancer cells
  by amplified spectra
  Raman scattering

- There are now 3 types of particles in use.
  It would be great if it only took 1

## Formalized ML task

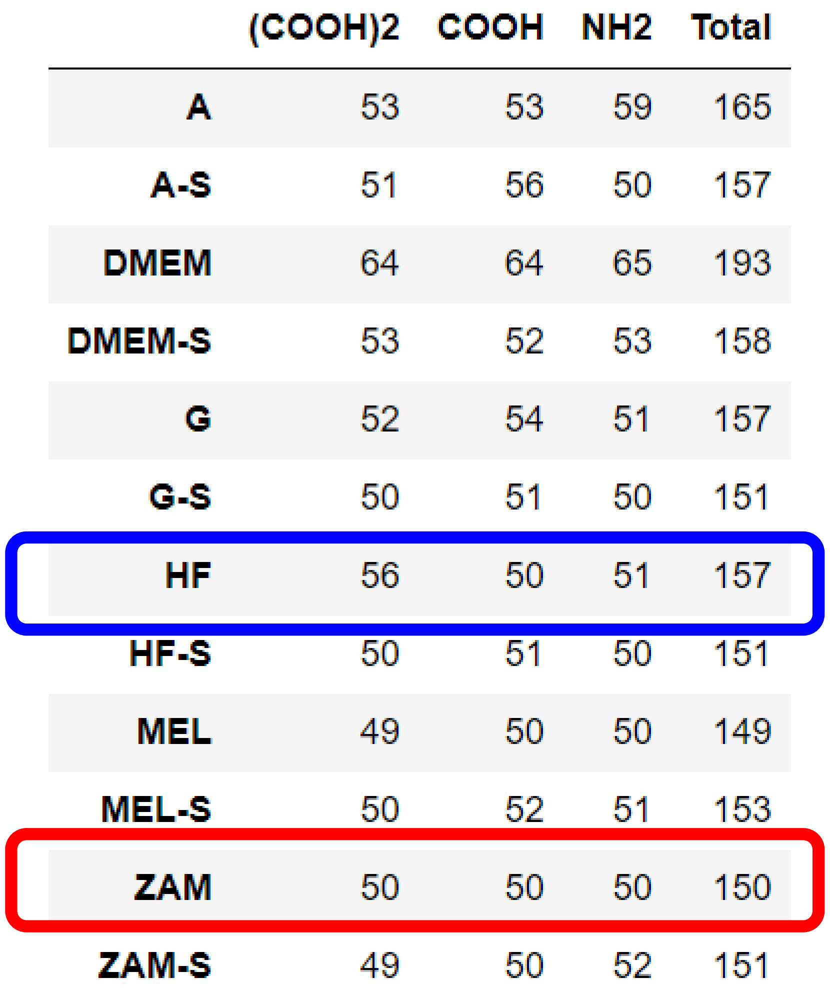
- **Classification**

- **Exploring the quality of the model depending on the input**

# Data

## Particle Coverage

Cell types



|  | (COOH)2 | COOH | NH2 | Total |
|---|---|---|---|---|
| **A** | 53 | 53 | 59 | 165 |
| **A-S** | 51 | 56 | 50 | 157 |
| **DMEM** | 64 | 64 | 65 | 193 |
| **DMEM-S** | 53 | 52 | 53 | 158 |
| **G** | 52 | 54 | 51 | 157 |
| **G-S** | 50 | 51 | 50 | 151 |
| **HF** | 56 | 50 | 51 | 157 |
| **HF-S** | 50 | 51 | 50 | 151 |
| **MEL** | 49 | 50 | 50 | 149 |
| **MEL-S** | 50 | 52 | 51 | 153 |
| **ZAM** | 50 | 50 | 50 | 150 |
| **ZAM-S** | 49 | 50 | 52 | 151 |

**Healthy cells**

**Cancer cells**

*https://www.kaggle.com/datasets/andriitrelin/cells-raman-spectra*

*Erzina et al. Sensors & Actuators: B. Chemical 308 (2020) 127660*

# Data

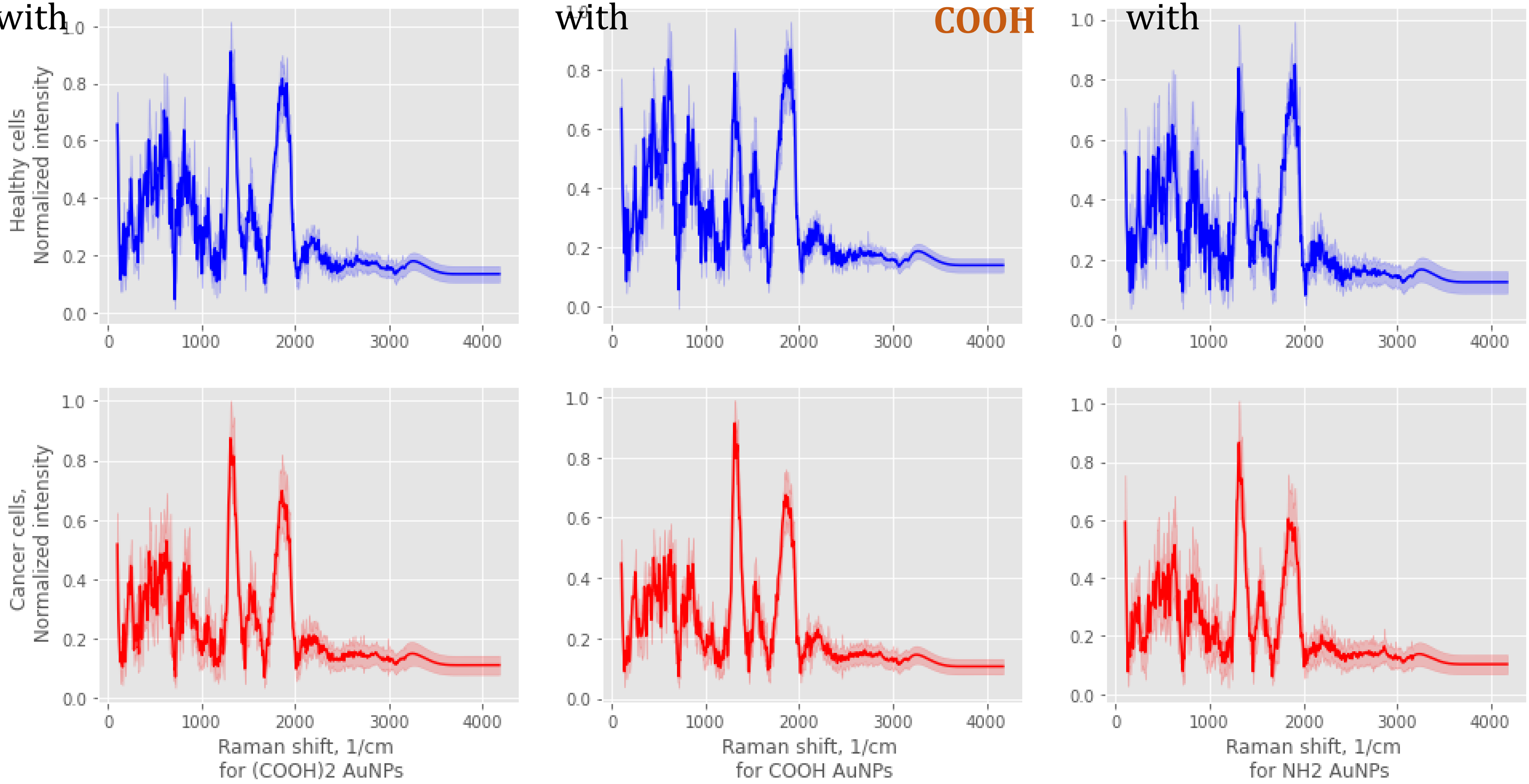
Particle amplification with -(COOH)2

Particle amplification with -COOH

Particle amplification with -NH2

Healthy cells

Cancer cells



In every spectrum:

**2000**

**features**

Total per sample:

**6000**

**features**

# Contents

- Introduction to the subject
- Problem statement
- **Data research**
- Applying machine learning techniques
  - Dimensionality reduction
  - Classification
- Task modification

# EDA and cleaning

- Data gaps: no

- Outliers: none

# Statistical analysis

- Let's look for spectral components that are different in healthy and sick normalized spectra, aiming to build a model/baseline without using ML.

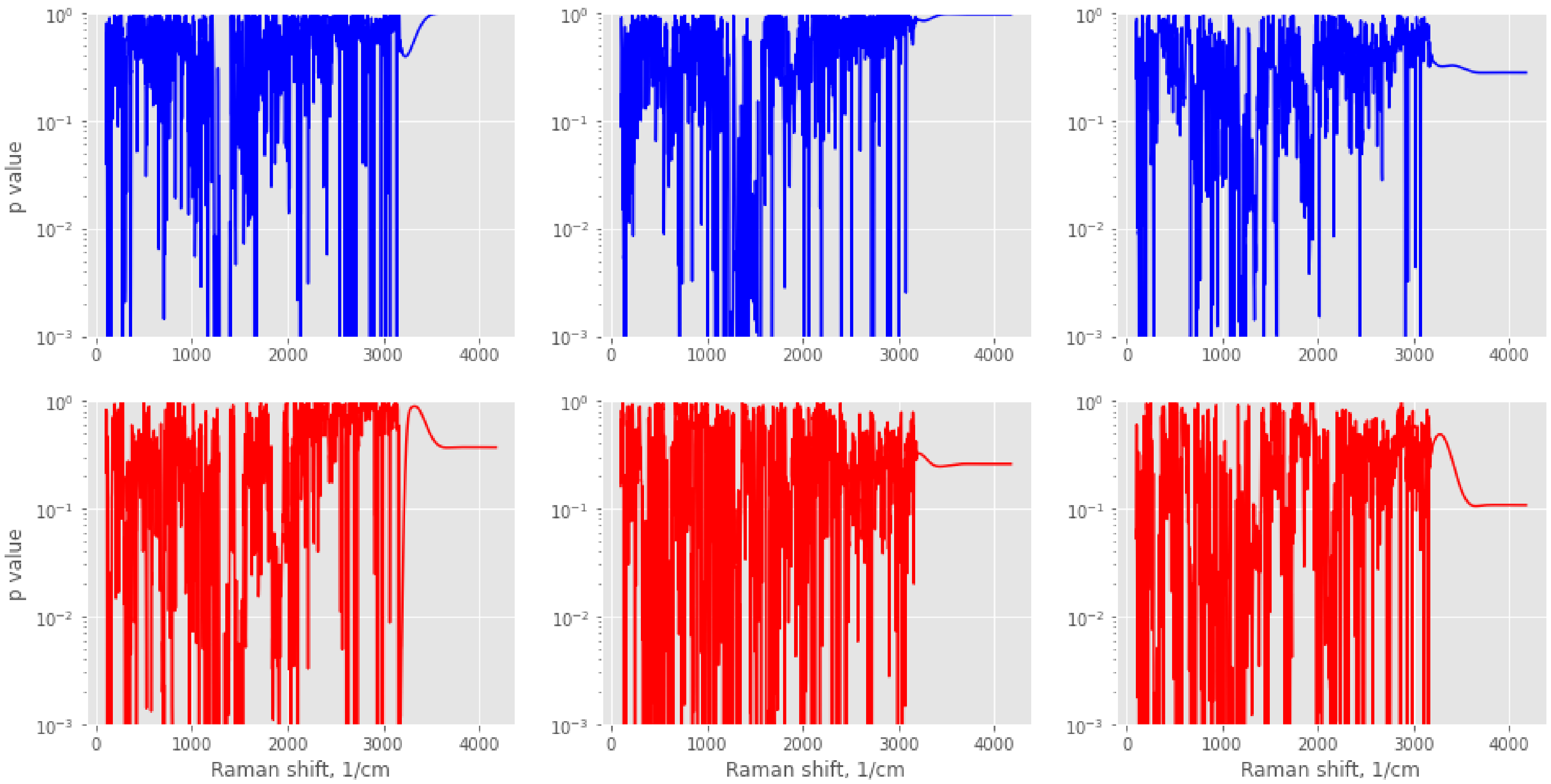
Expectation:

1) **Check the distribution in each trait for normality**
2) **Choose the type of test**
3) **Test**

# Statistical analysis

- Let's look for spectral components that are different in healthy and sick normalized spectra.
  1) Let's check the distribution in each trait for normality

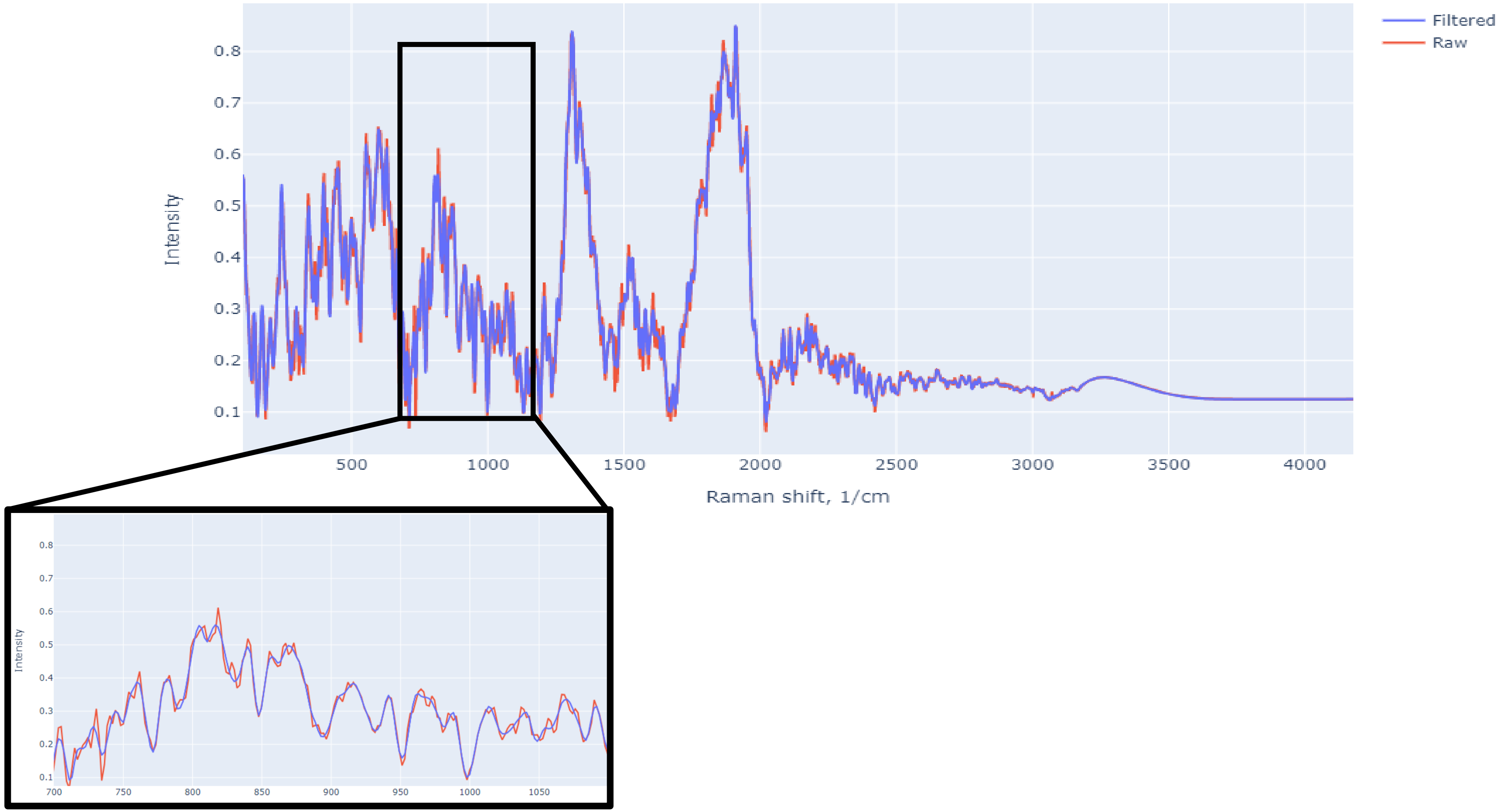
Probability that intensities of spectral components are normally distributed



# Savitsky-Golay filtering

Trade-off between smoothing and informativity, as some peaks also may be sharp

Mean spectrum of filered HF NH2 spectra



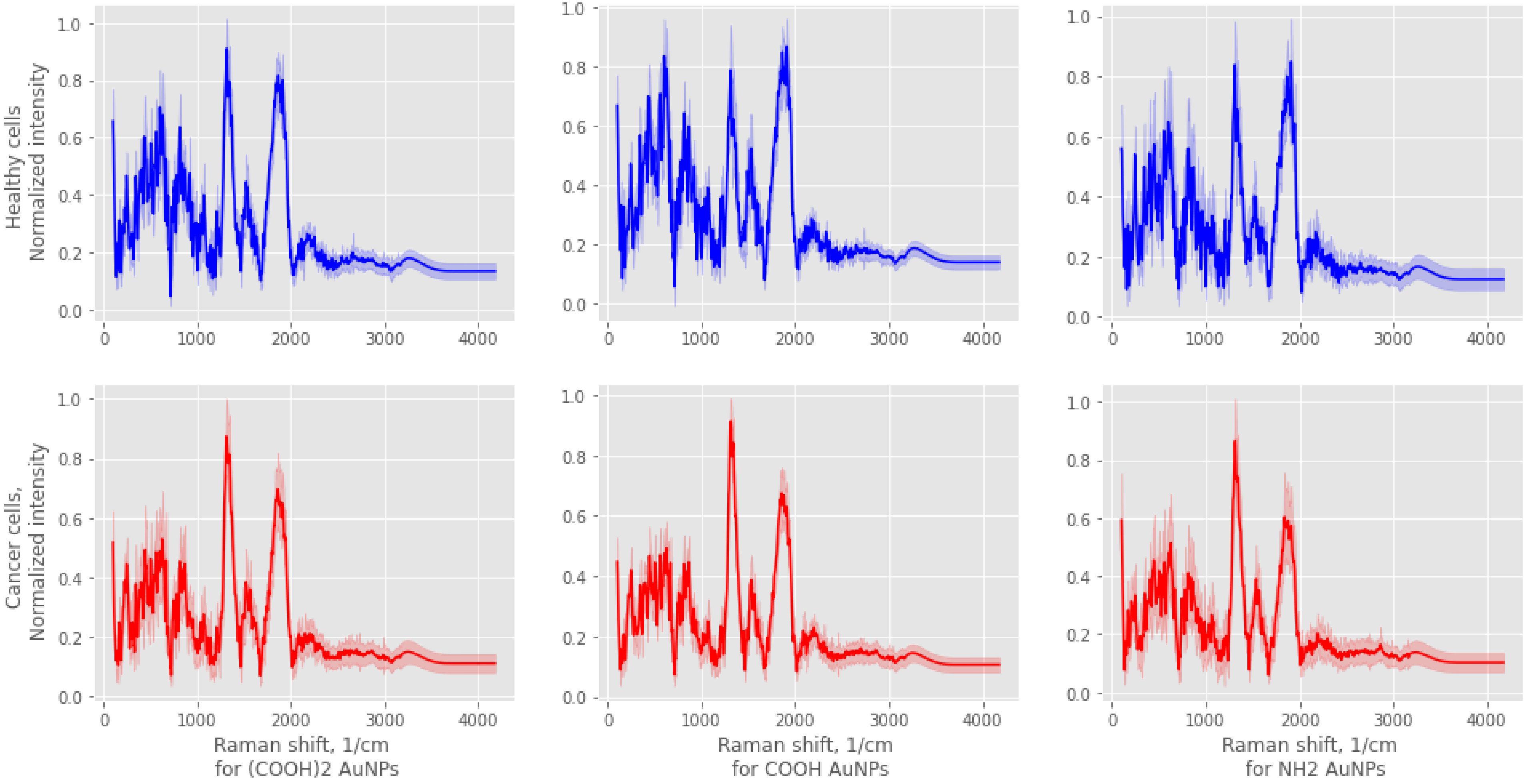
# Savitsky-Golay filter result

All spectra studied, mean+-std:

Raman spectra, mean+-std, Savitzky-Golay filtered

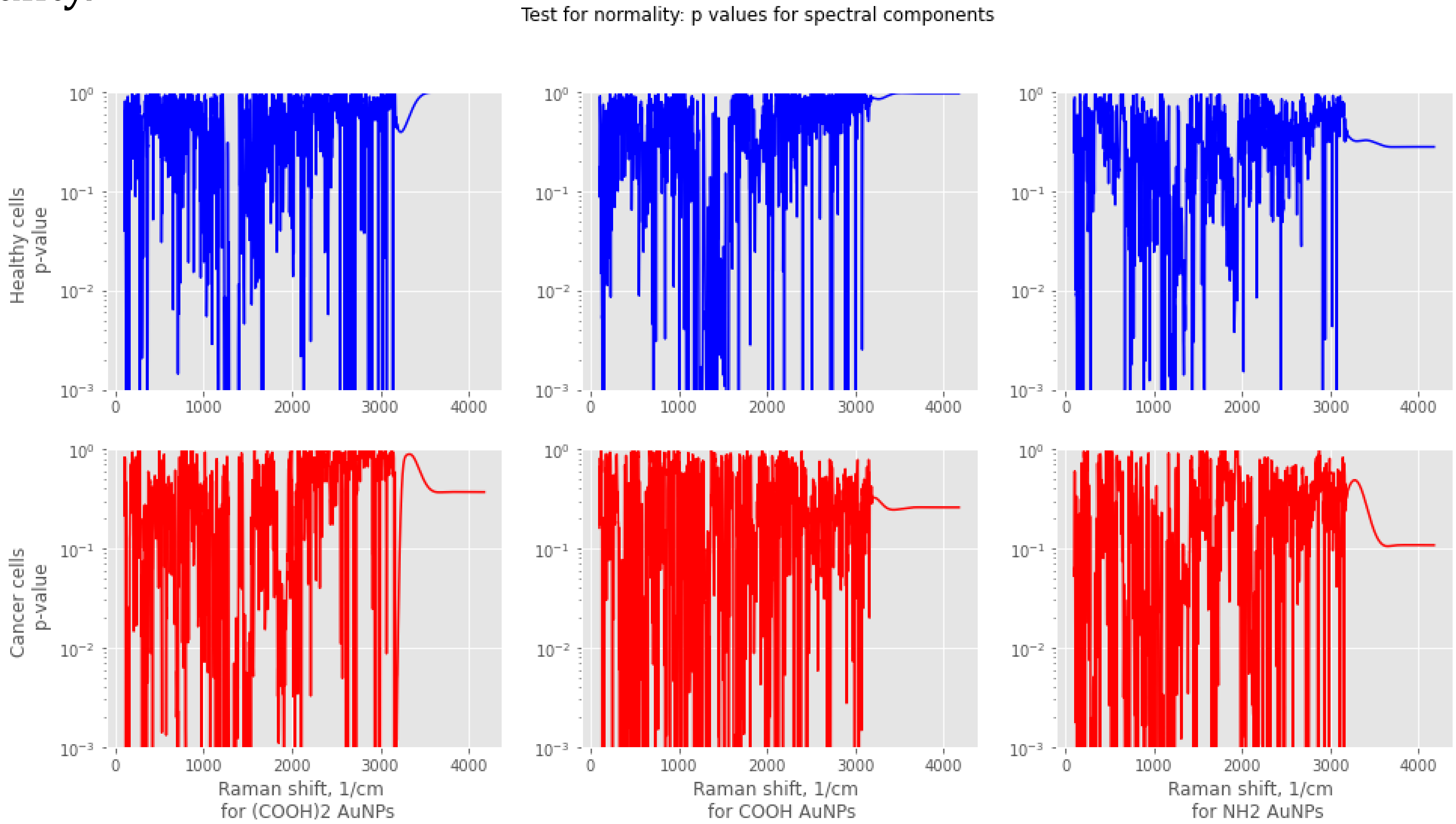
**Healthy cells**

**Cancer cells**



# Statistical analysis

After the Savitsky-Golay filter, again check the distribution in each trait for normality.

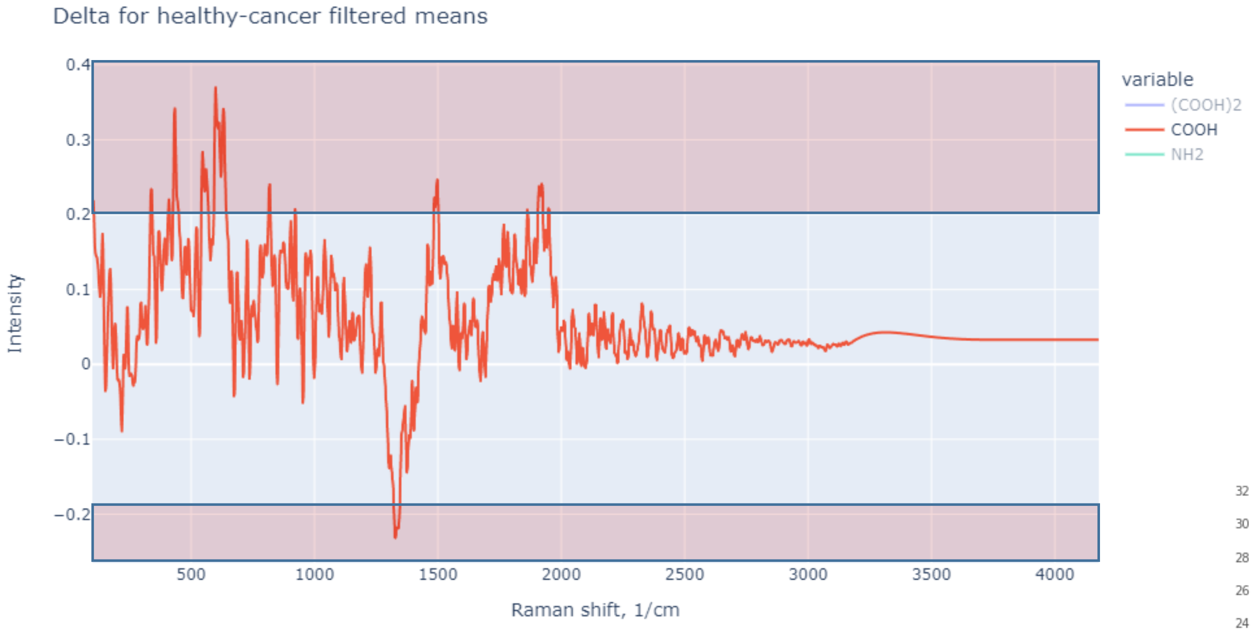


Again, very different p-values.

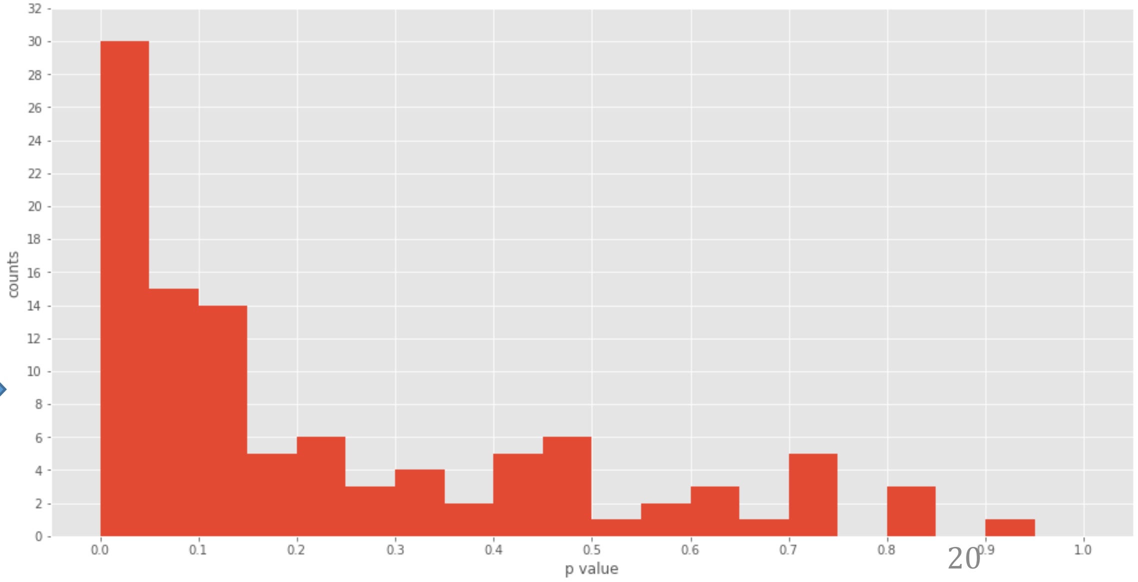
But we don't need all of them, we only need them in spades

# Looking for regions of interest

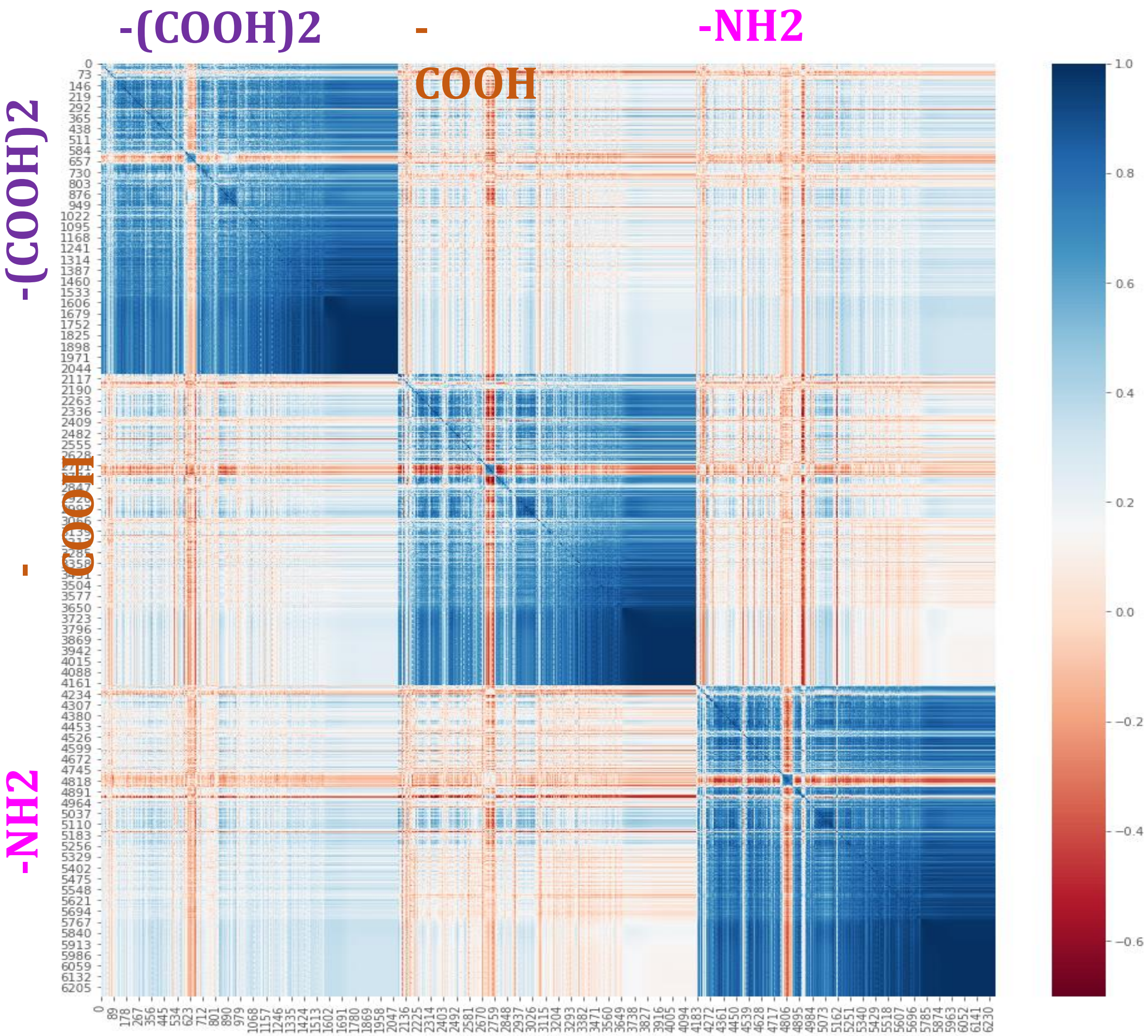
- Delta between mean spectra:



- Let's select those components where abs(delta) > ( For UND there are 106 of them.
- Let's check them for normality and get the followin histogram:
- We see that the p-value<0.05 for 30 points out of 106.



# Correlation matrix



- We use **Spearman correlation**, since we already know that not all signs have a normal distribution
- We see: the data in one spectrum are correlated with each other
- Conclusion: it is worth trying **downsizing**

# Contents

- Introduction to the subject
- Problem statement
- Data research
- Applying machine learning techniques
    - **Dimensionality reduction**
    - Classification
- Task modification

# Dimensionality reduction: PCA

Thousands of signs

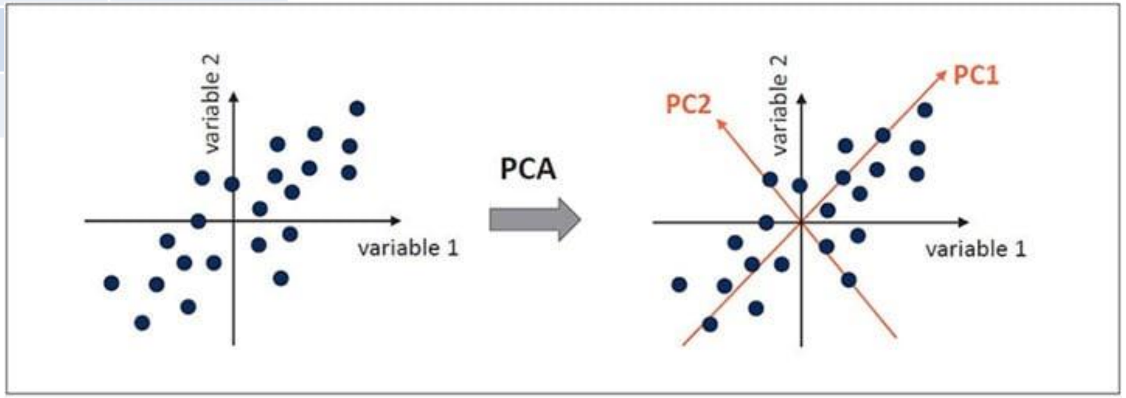
The data is very large!

| Wave-number | Sample 1 Intensity | Sample 2 Intensity |
|---|---|---|
| 600 | | |
| 601 | | |
| 602 | | |
| 603 | | |
| . | | |
| . | | |
| . | | |
| . | | |
| . | | |
| . | | |
| . | | |
| . | | |

| Principal component | Sample 1 Intensity | Sample 2 Intensity |
|---|---|---|
| 1 | | |
| 2 | | |
| 3 | | |
| 4 | | |
| 5 | | |

Just a few components make a difference!

=> Dimensionality much less

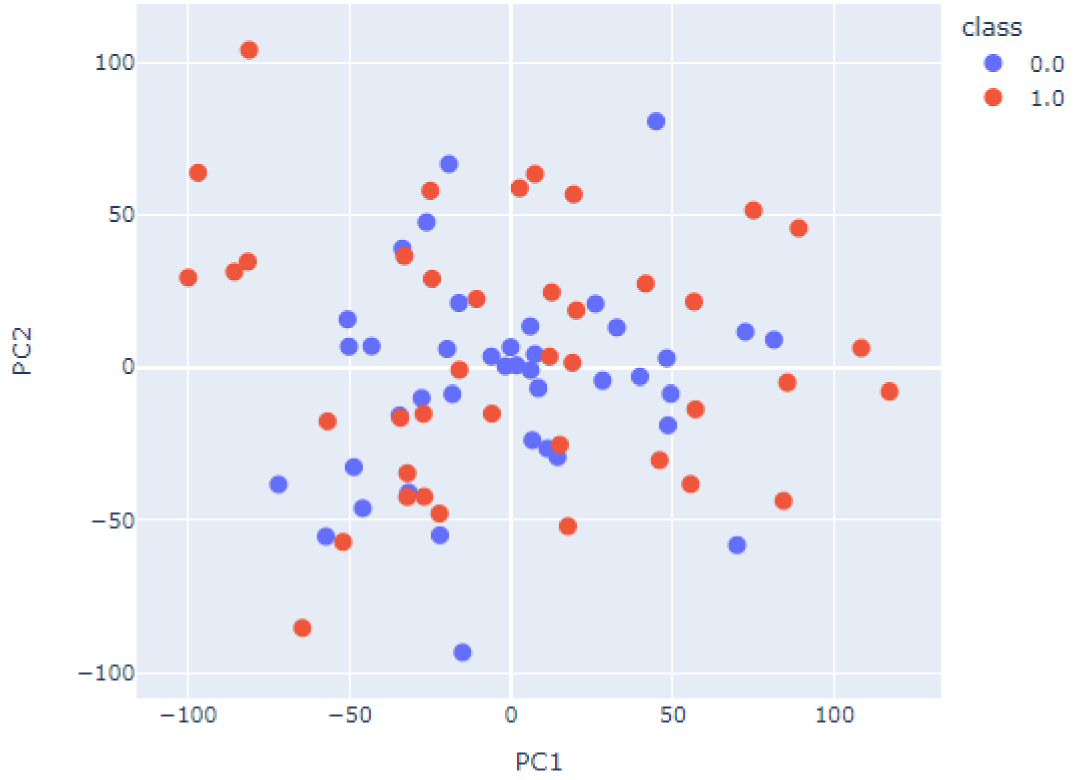


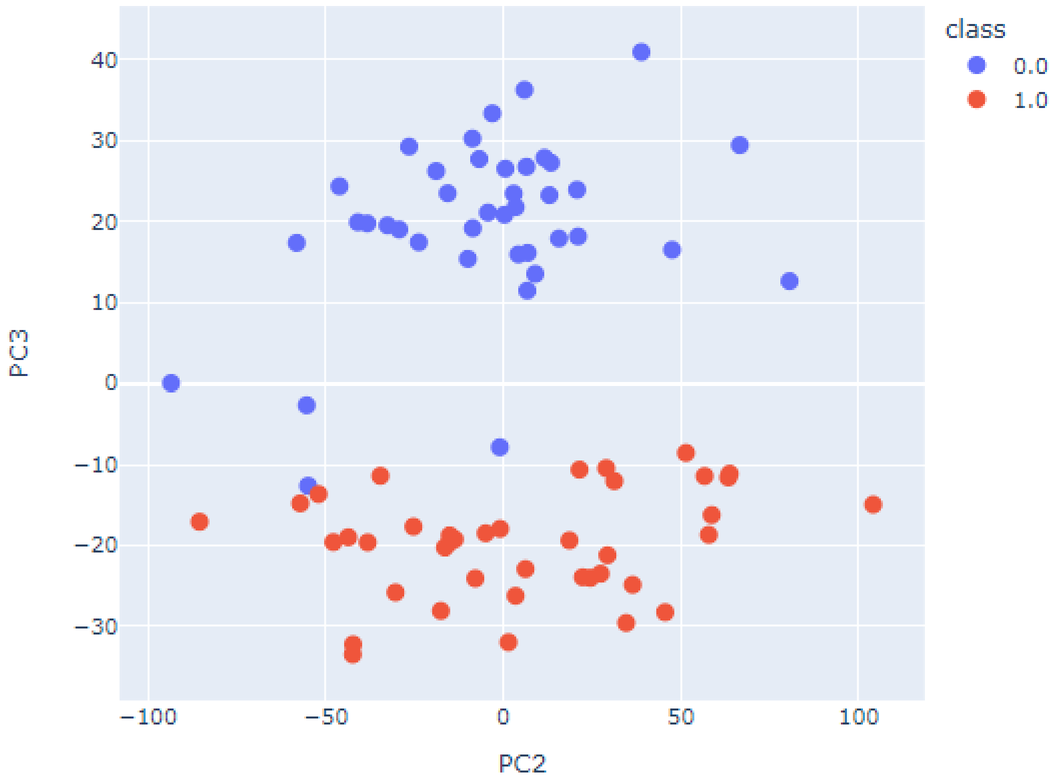
Source: ourcodingclub.github.io

# PCA - Principal Component Analysis

Pre-processing:

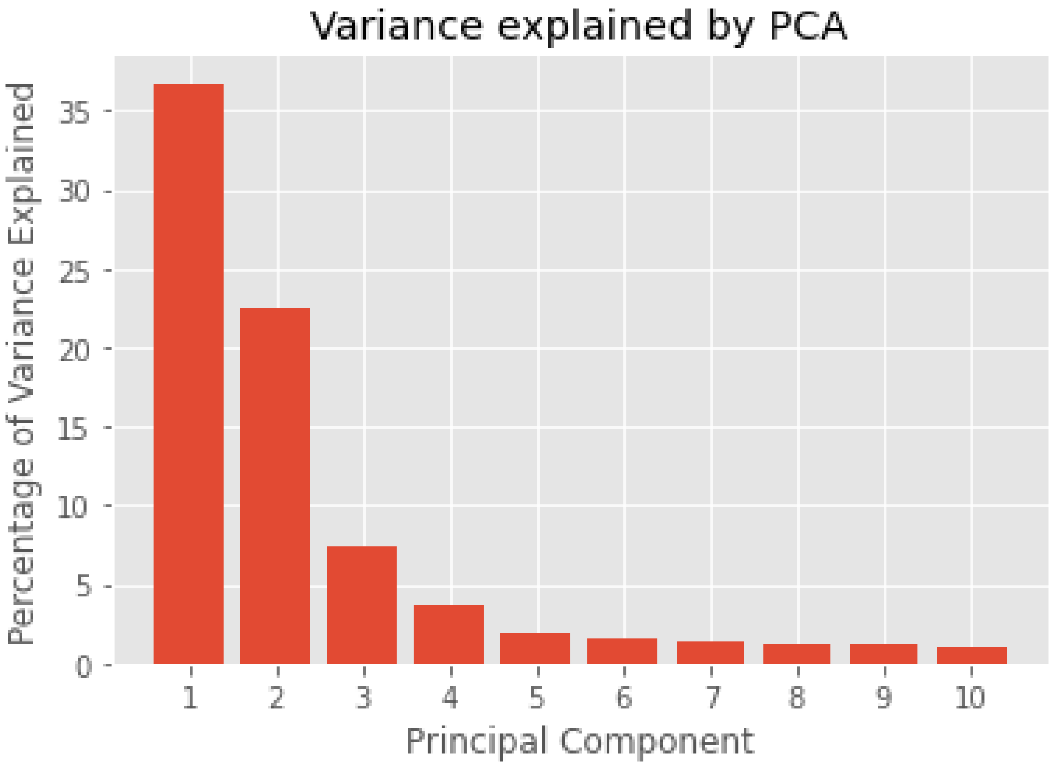
- Each spectrum was normalized by the area under its curve
- Then, each trait was standardized across all spectra

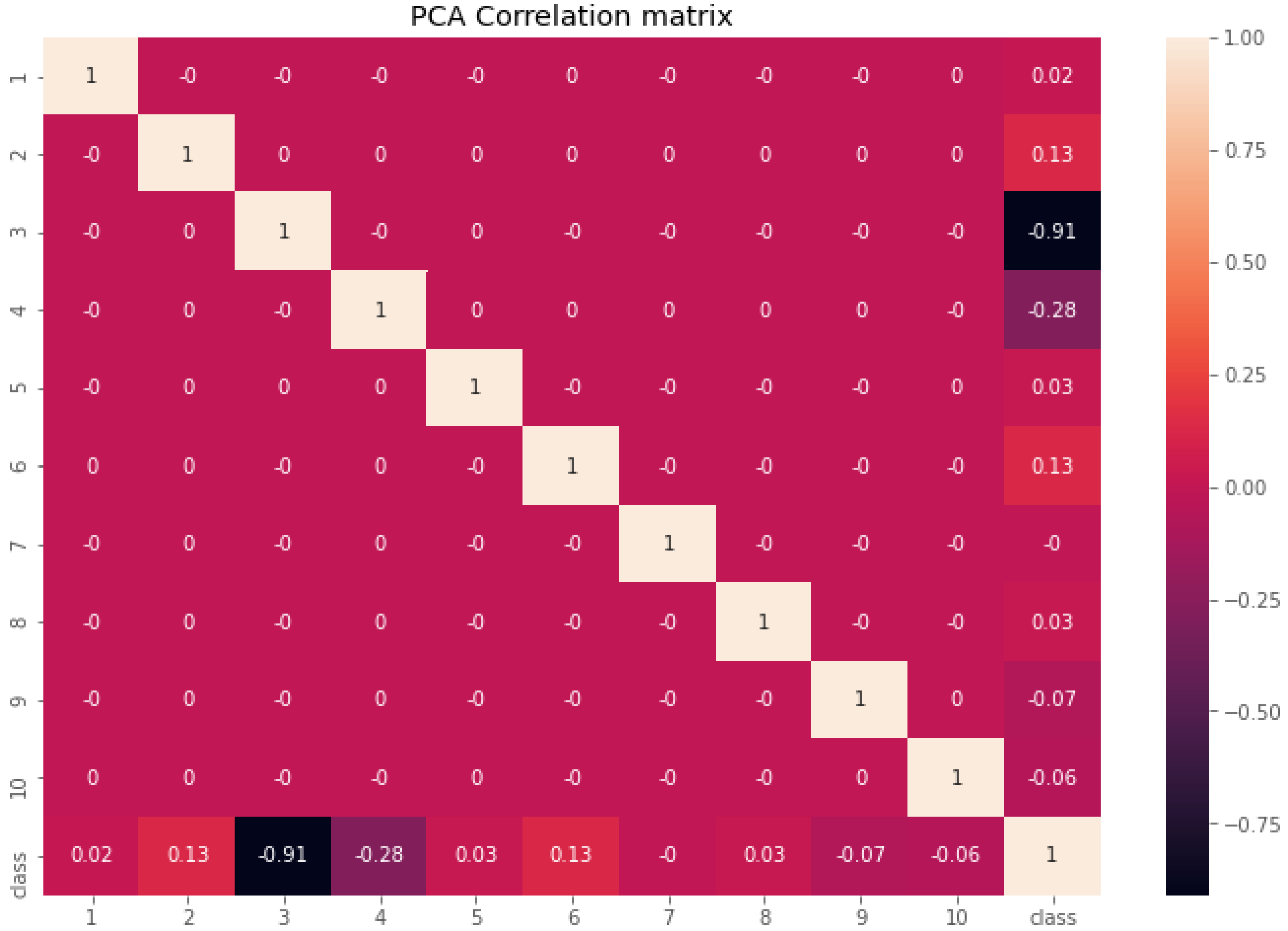




On PC3 we can observe almost complete separation of classes

# PCA analytics





- all PCA components are orthogonal to each other, their mutual correlations = 0
- PC3 is strongly correlated with the class label

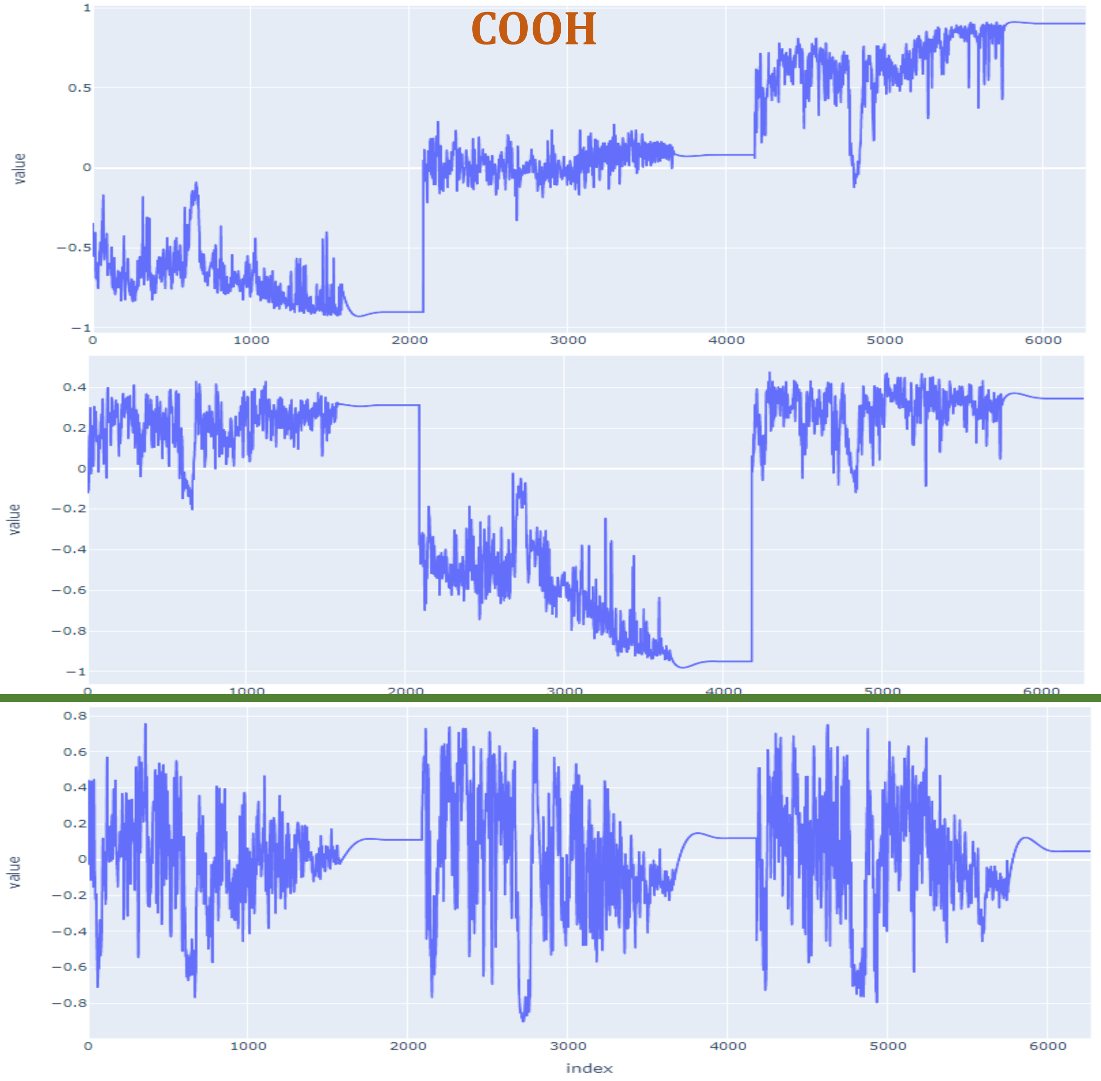
# PCA loadings

-(COOH)2 -COOH -NH2

PC 1

PC 2

PC 3

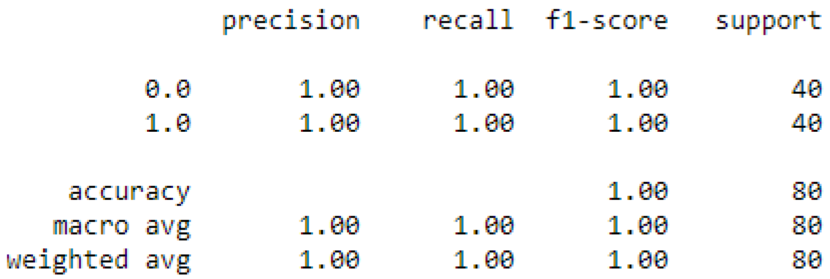


# Contents

- Introduction to the subject
- Problem statement
- Data research
- Applying machine learning techniques
  - Dimensionality reduction
  - **Classification**
- Task modification

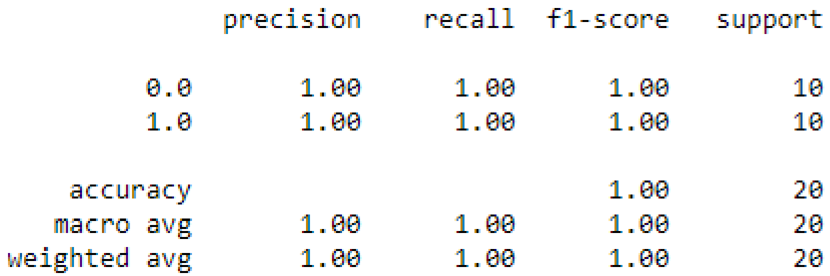
# Logistic Regression

Train set



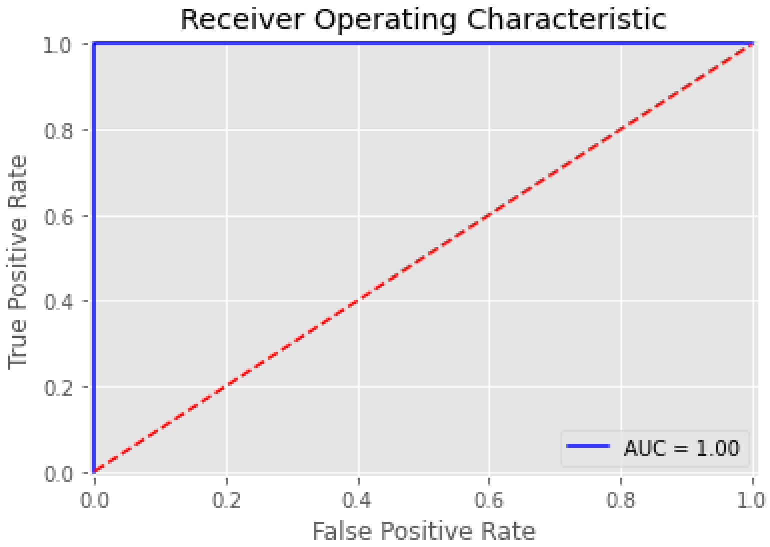
```
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        40
         1.0       1.00      1.00      1.00        40

    accuracy                           1.00        80
   macro avg       1.00      1.00      1.00        80
weighted avg       1.00      1.00      1.00        80
```

Test set

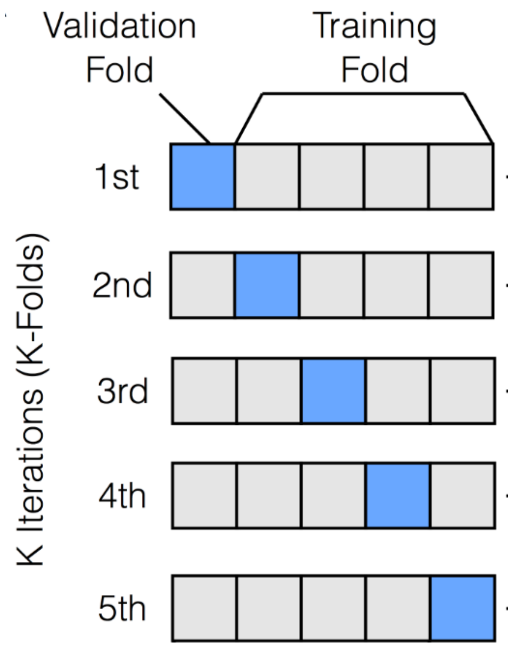


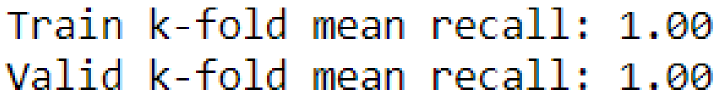
```
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        10
         1.0       1.00      1.00      1.00        10

    accuracy                           1.00        20
   macro avg       1.00      1.00      1.00        20
weighted avg       1.00      1.00      1.00        20
```

100% Quality even on the test set 🧐

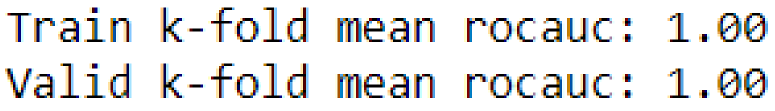


# Cross Validation





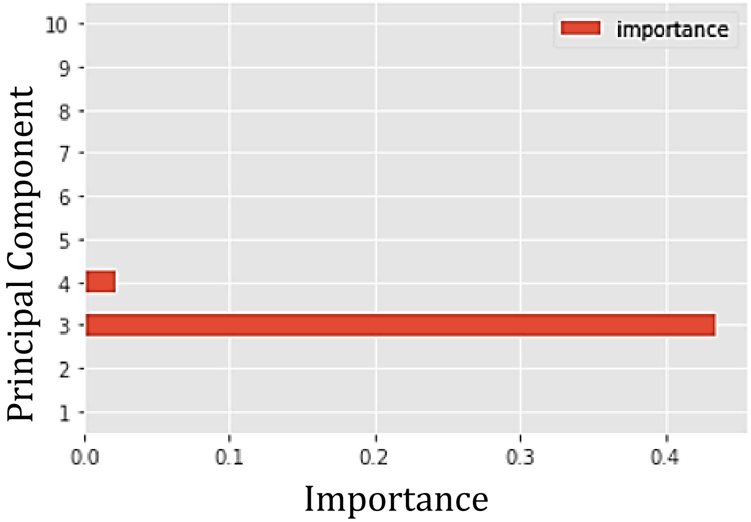
```
Train k-fold mean recall: 1.00
Valid k-fold mean recall: 1.00
```

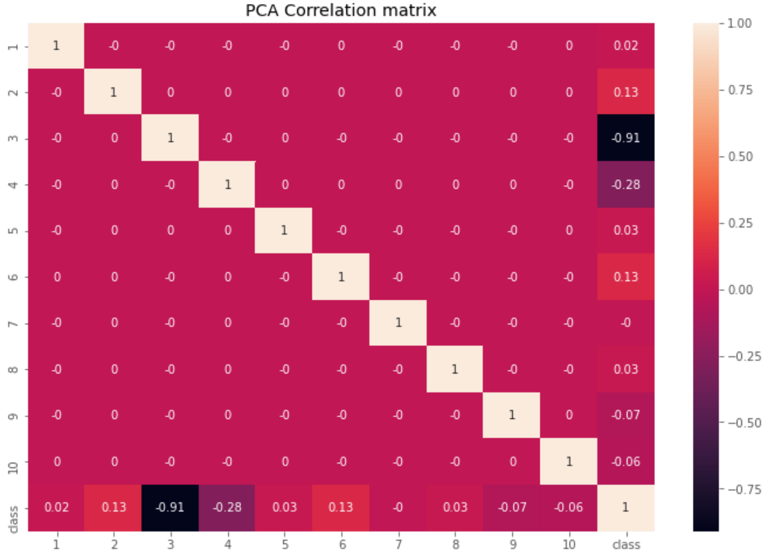


```
Train k-fold mean rocauc: 1.00
Valid k-fold mean rocauc: 1.00
```

Still 100% quality 🧐 🧐

# Feature importances




PCA Correlation matrix

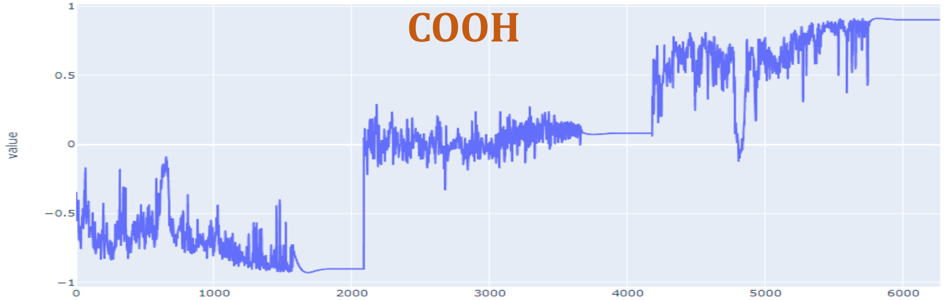
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | class |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | -0 | -0 | -0 | -0 | 0 | -0 | -0 | -0 | 0 | 0.02 |
| 2 | -0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.13 |
| 3 | -0 | 0 | 1 | -0 | 0 | -0 | -0 | -0 | -0 | -0 | -0.91 |
| 4 | -0 | 0 | -0 | 1 | 0 | 0 | 0 | 0 | 0 | -0 | -0.28 |
| 5 | -0 | 0 | 0 | 0 | 1 | -0 | -0 | -0 | -0 | 0 | 0.03 |
| 6 | 0 | 0 | -0 | 0 | -0 | 1 | -0 | -0 | -0 | -0 | 0.13 |
| 7 | -0 | 0 | -0 | 0 | -0 | -0 | 1 | -0 | -0 | -0 | -0 |
| 8 | -0 | 0 | -0 | 0 | -0 | -0 | -0 | 1 | -0 | -0 | 0.03 |
| 9 | -0 | 0 | -0 | 0 | -0 | -0 | -0 | -0 | 1 | 0 | -0.07 |
| 10 | 0 | 0 | -0 | -0 | 0 | -0 | -0 | -0 | 0 | 1 | -0.06 |
| class | 0.02 | 0.13 | -0.91 | -0.28 | 0.03 | 0.13 | -0 | 0.03 | -0.07 | -0.06 | 1 |

-(COOH)2

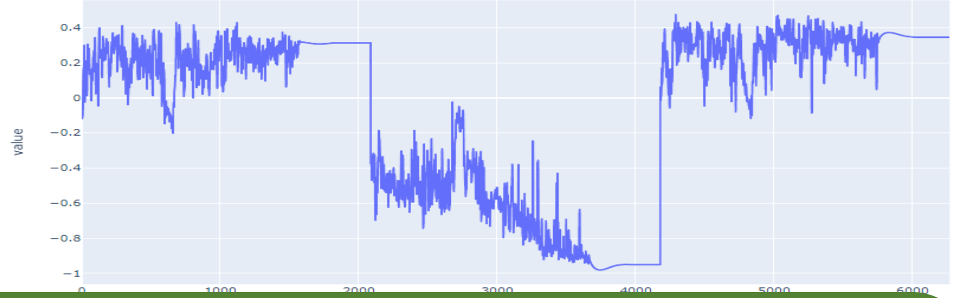
-COOH

-NH2

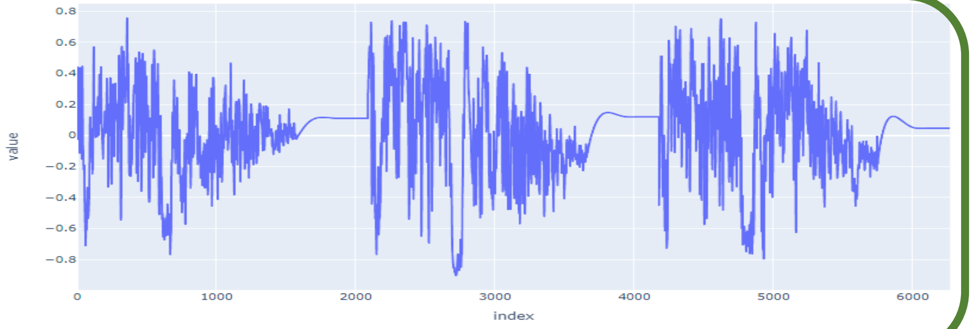
PC 1



PC 2



PC 3

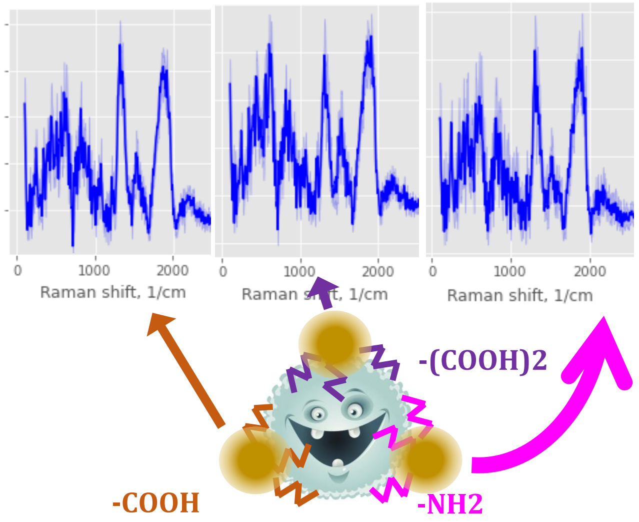


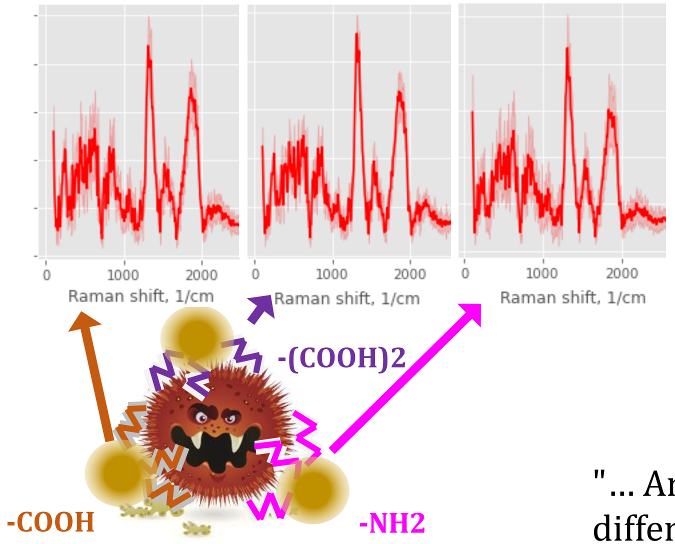
# Contents

- Introduction to the subject
- Problem statement
- Data research
- Applying machine learning techniques
  - Dimensionality reduction
  - Classification
- **Task modification**

# We use only 1 type of particles

**It was:**



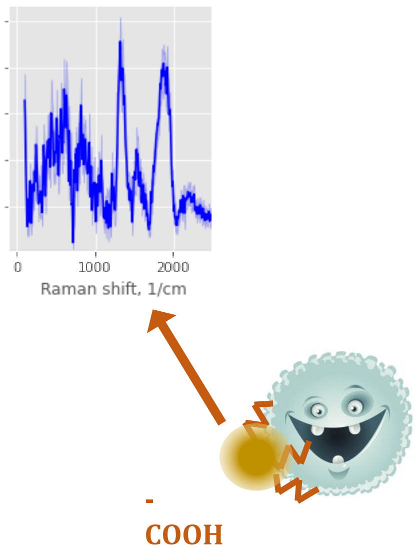


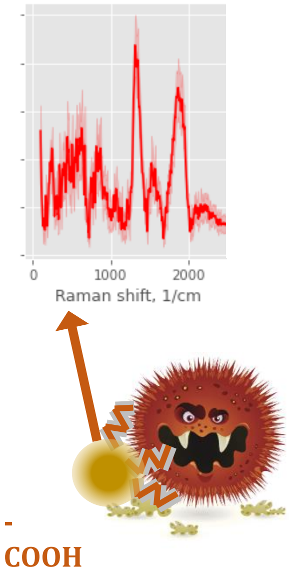
**6000**

features

"... And if there's no difference, why pay more?"

**Stal:**





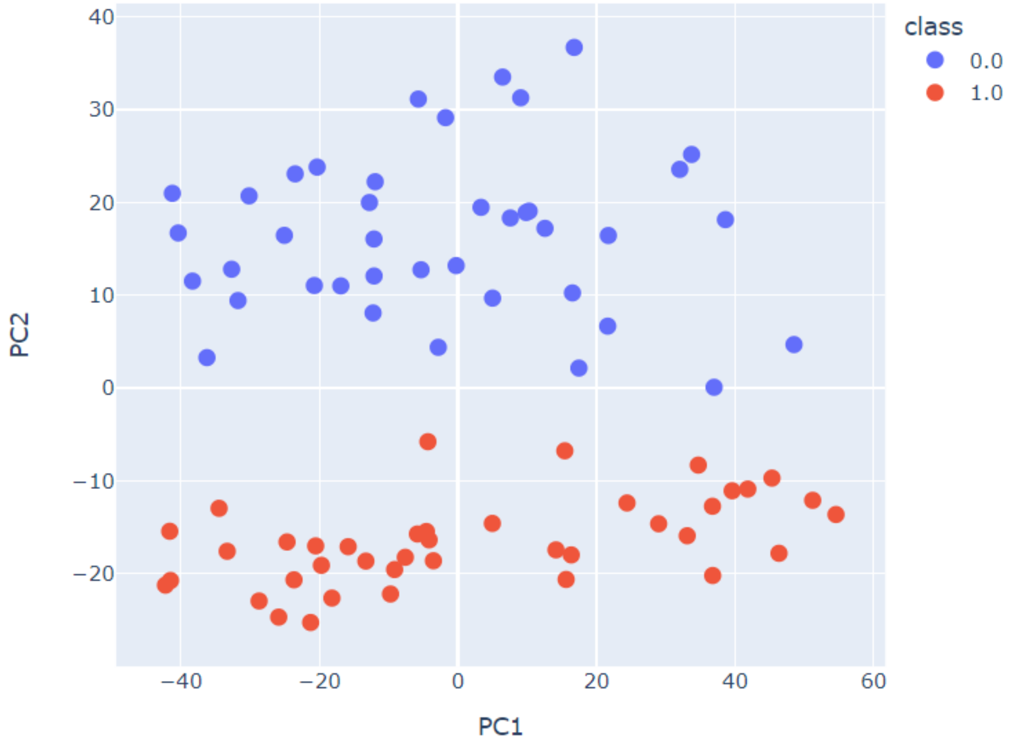
**2000**

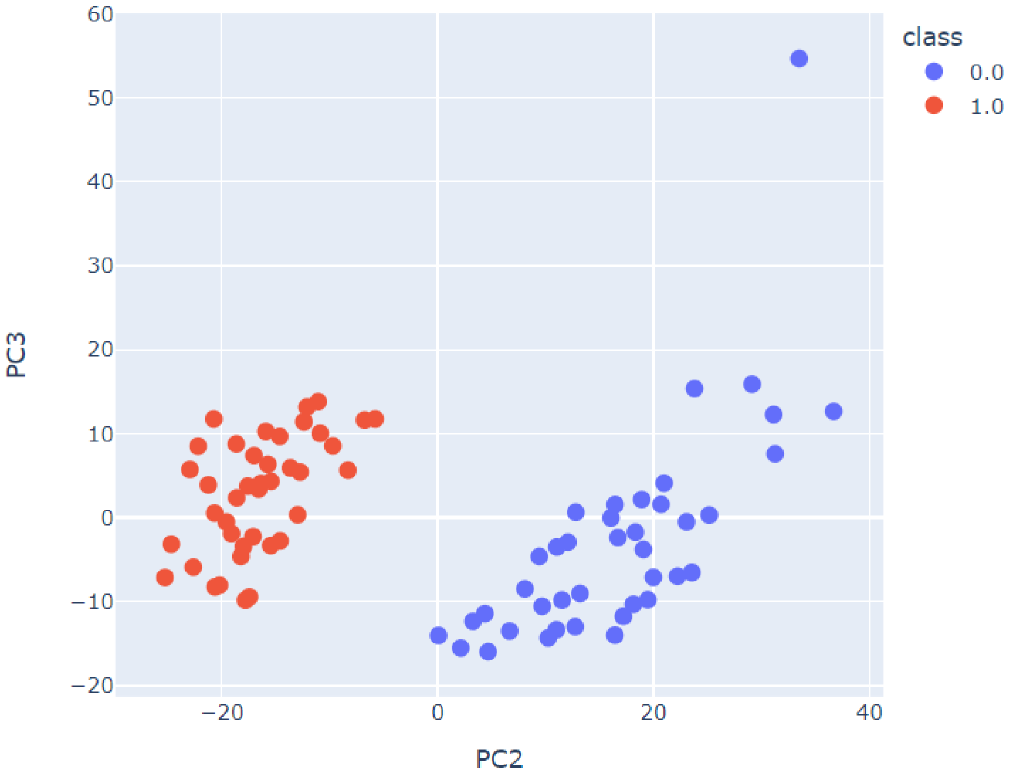
features

**3x less**

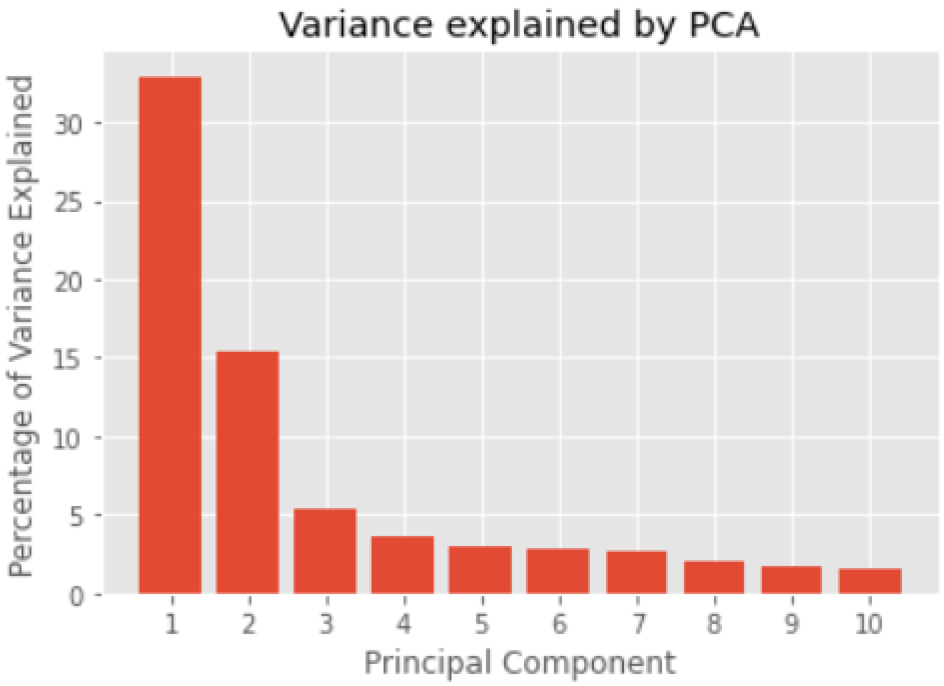
AuNPs and spectra

# PCA - Principal Component Analysis

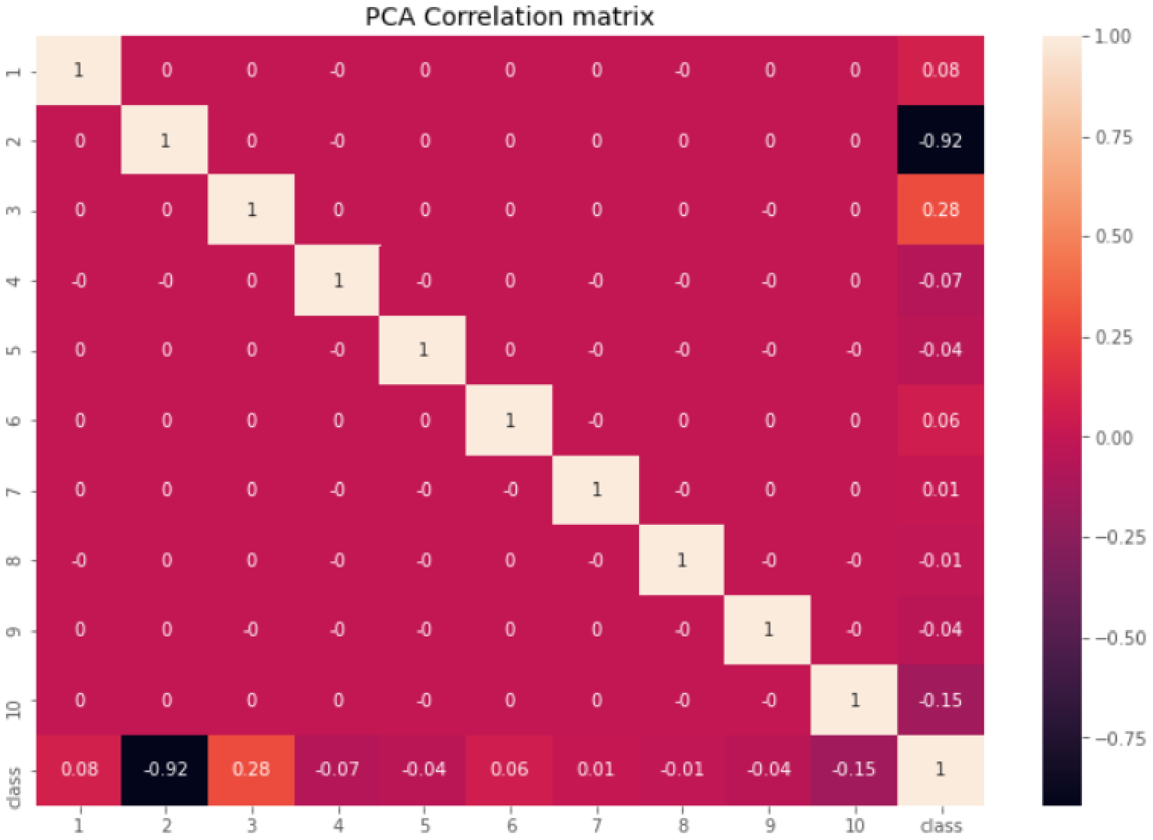




# PCA analytics



Mutual correlation matrix

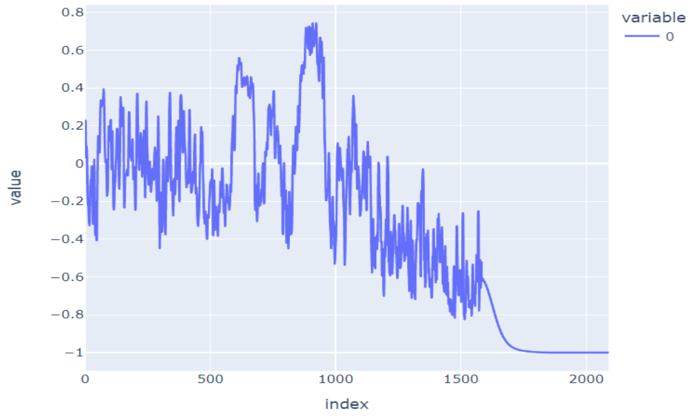


|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | class |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | -0 | 0 | 0 | 0 | -0 | 0 | 0 | 0.08 |
| 2 | 0 | 1 | 0 | -0 | 0 | 0 | 0 | 0 | 0 | 0 | -0.92 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | -0 | 0 | 0.28 |
| 4 | -0 | -0 | 0 | 1 | -0 | 0 | -0 | -0 | -0 | 0 | -0.07 |
| 5 | 0 | 0 | 0 | -0 | 1 | 0 | -0 | -0 | -0 | -0 | -0.04 |
| 6 | 0 | 0 | 0 | 0 | 0 | 1 | -0 | 0 | 0 | 0 | 0.06 |
| 7 | 0 | 0 | 0 | -0 | -0 | -0 | 1 | -0 | 0 | 0 | 0.01 |
| 8 | -0 | 0 | 0 | -0 | -0 | 0 | -0 | 1 | -0 | -0 | -0.01 |
| 9 | 0 | 0 | -0 | -0 | -0 | 0 | 0 | -0 | 1 | -0 | -0.04 |
| 10 | 0 | 0 | 0 | 0 | -0 | 0 | 0 | -0 | -0 | 1 | -0.15 |
| class | 0.08 | -0.92 | 0.28 | -0.07 | -0.04 | 0.06 | 0.01 | -0.01 | -0.04 | -0.15 | 1 |

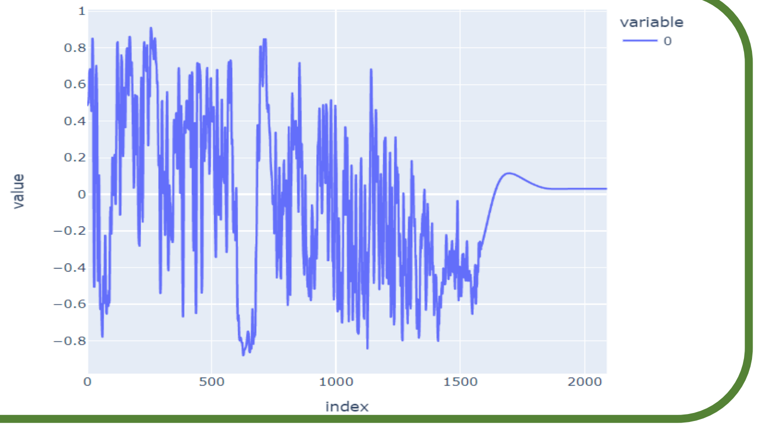
- all PCA components are orthogonal to each other, their mutual correlations = 0
- PC2 is strongly correlated with the class label

# PCA loadings

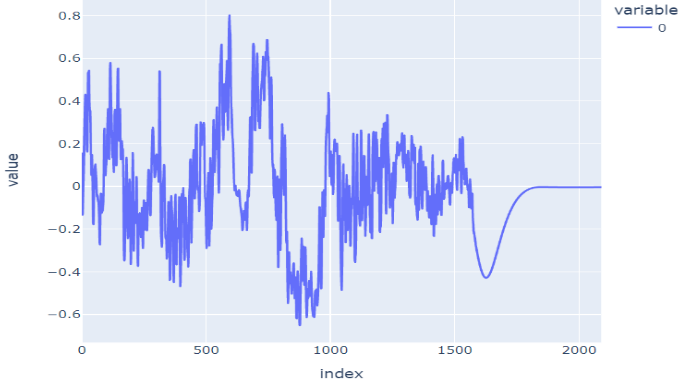
**PC 1**



**PC 2**

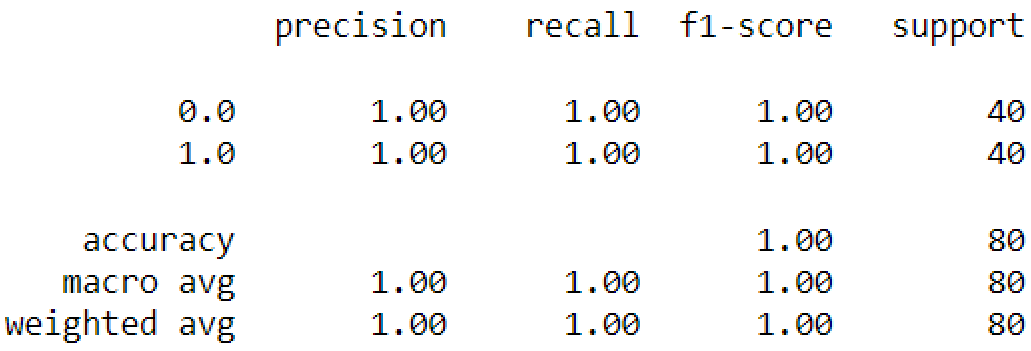


**PC 3**



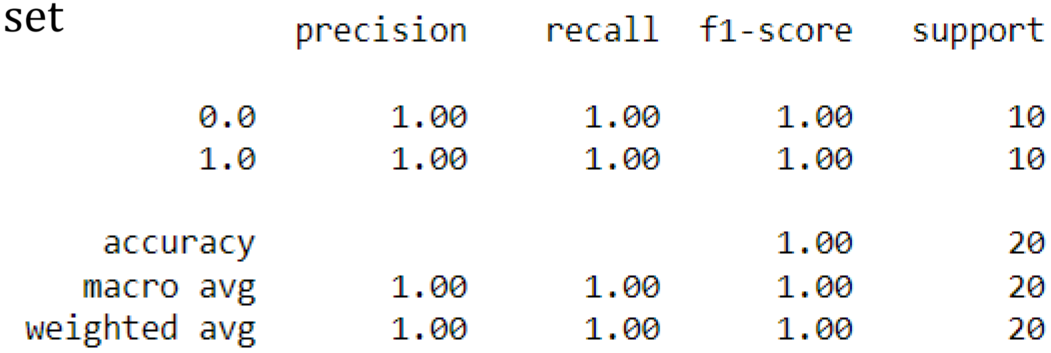
# Logistic Regression

Train set



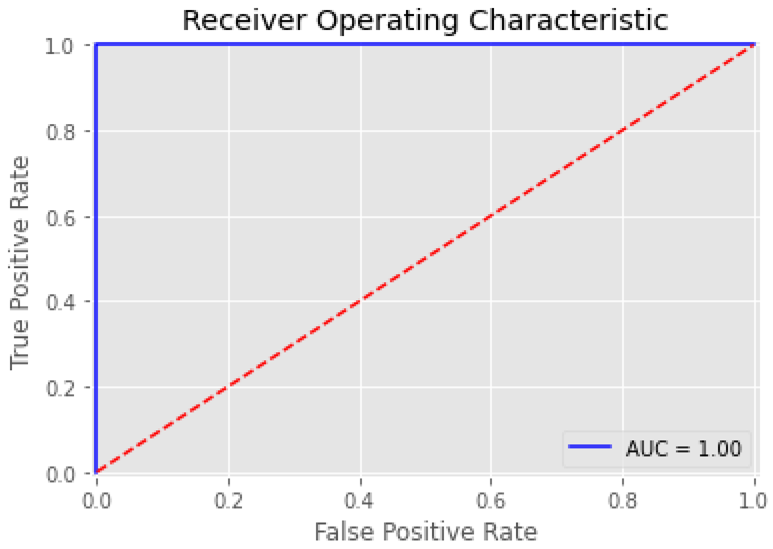
```
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        40
         1.0       1.00      1.00      1.00        40

    accuracy                           1.00        80
   macro avg       1.00      1.00      1.00        80
weighted avg       1.00      1.00      1.00        80
```

Test set

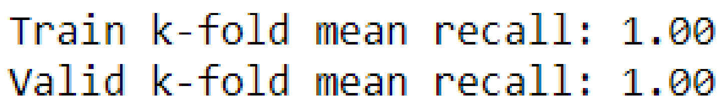


```
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        10
         1.0       1.00      1.00      1.00        10

    accuracy                           1.00        20
   macro avg       1.00      1.00      1.00        20
weighted avg       1.00      1.00      1.00        20
```

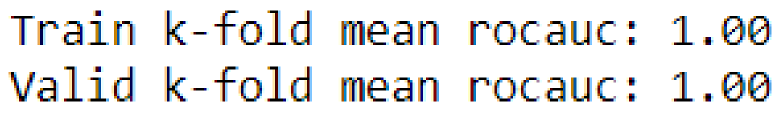
100% Quality even on the test set 🧐 🧐 🧐



# Cross Validation



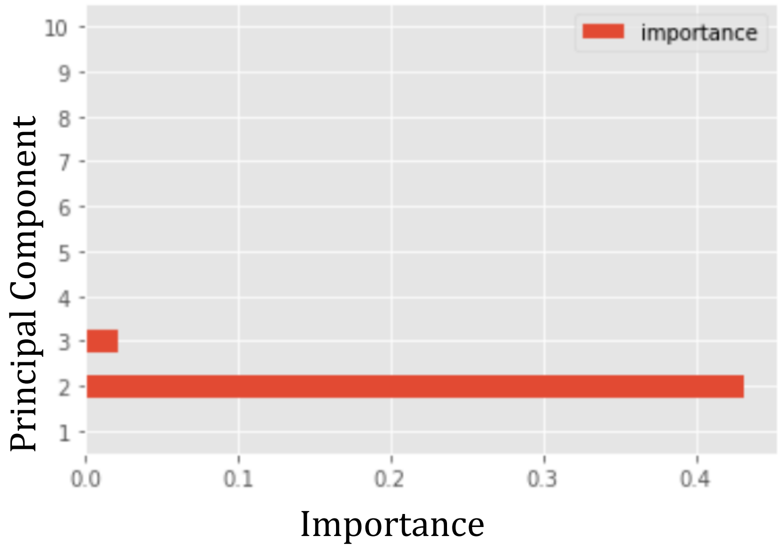
```
Train k-fold mean recall: 1.00
Valid k-fold mean recall: 1.00
```

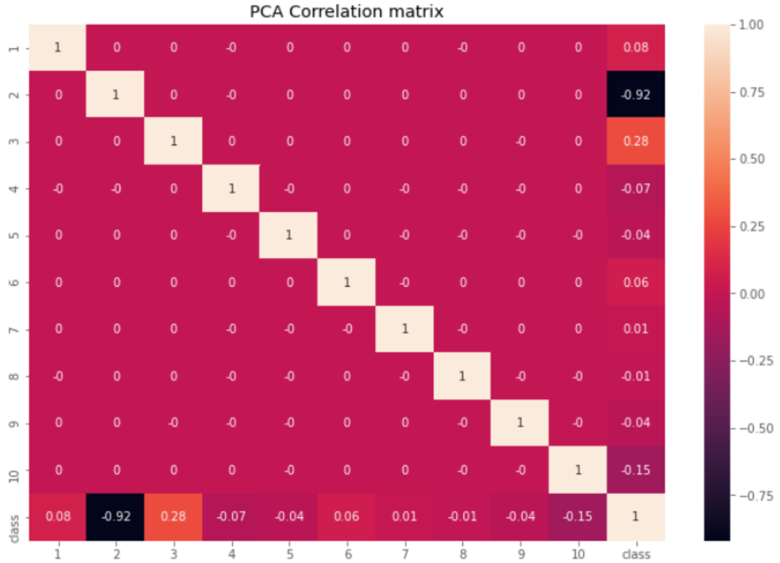


```
Train k-fold mean rocauc: 1.00
Valid k-fold mean rocauc: 1.00
```

Still 100% quality 🧐 🧐 🧐 🧐

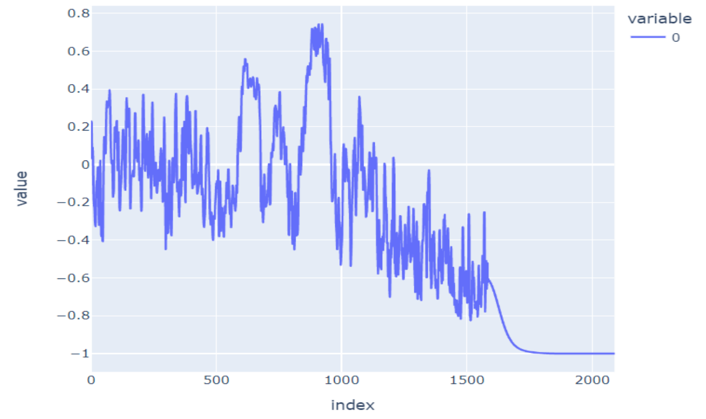
# Feature importances



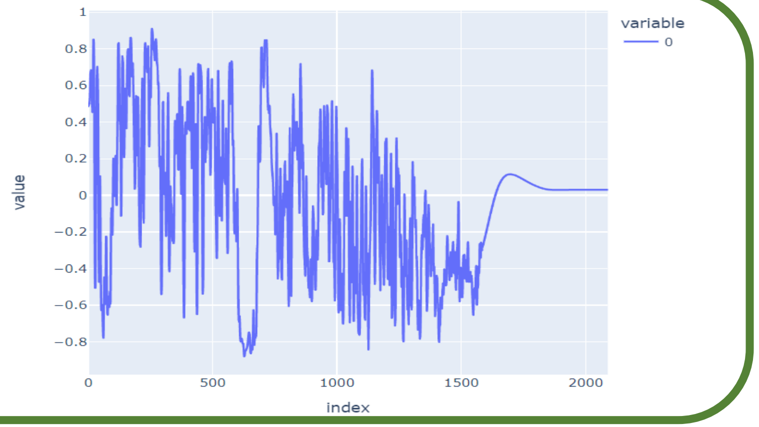

PCA Correlation matrix

|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | class |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | -0 | 0 | 0 | 0 | -0 | 0 | 0 | 0.08 |
| 2 | 0 | 1 | 0 | -0 | 0 | 0 | 0 | 0 | 0 | 0 | -0.92 |
| 3 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | -0 | 0 | 0.28 |
| 4 | -0 | -0 | 0 | 1 | -0 | 0 | -0 | -0 | -0 | 0 | -0.07 |
| 5 | 0 | 0 | 0 | -0 | 1 | 0 | -0 | -0 | -0 | -0 | -0.04 |
| 6 | 0 | 0 | 0 | 0 | 0 | 1 | -0 | 0 | 0 | 0 | 0.06 |
| 7 | 0 | 0 | 0 | -0 | -0 | -0 | 1 | -0 | 0 | 0 | 0.01 |
| 8 | -0 | 0 | 0 | -0 | -0 | 0 | -0 | 1 | -0 | -0 | -0.01 |
| 9 | 0 | 0 | -0 | -0 | -0 | 0 | 0 | -0 | 1 | -0 | -0.04 |
| 10 | 0 | 0 | 0 | 0 | -0 | 0 | 0 | -0 | -0 | 1 | -0.15 |
| class | 0.08 | -0.92 | 0.28 | -0.07 | -0.04 | 0.06 | 0.01 | -0.01 | -0.04 | -0.15 | 1 |

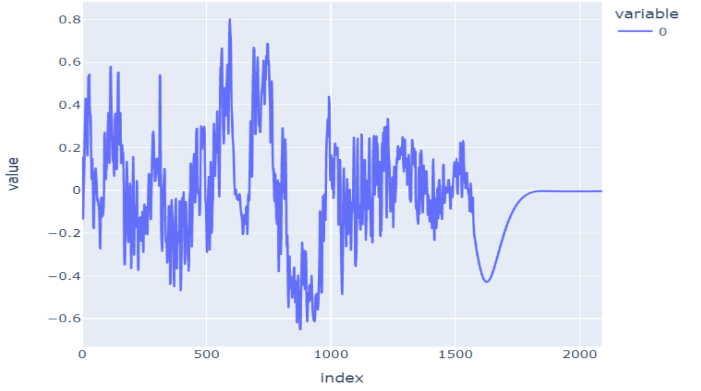
PC 1



PC 2



PC 3



# Conclusion

- Even with the use of a single particle type, it is quite easy to implement cell classification.
- To check the "suspicious perfection" of the results:
  - The data was examined for leaks - no leaks were found
  - Cross-validation was performed - quality is still 100%

- ToDo:
  - More measurements

# The end

Do you have any questions,

feedback or want to stay in touch?

Please contact me:

✉ a.merdalimova@gmail.com
linkedin.com/in/anastasiia-merdalimova/
github.com/Asya23/

Anastasiia Merdalimova

February 2023