# New Cytomine modules for user behavior analytics in digital pathology

Master Thesis

Academic year 2017-2018
University of Liège - Faculty of Applied Sciences


Graduation Studies conducted for obtaining the Master's degree in
Computer Sciences by Laurent Vanhee

June 2018

# Contents

# 1 Abstract

In the medical field, doctors and researchers need to be able to observe and interpret cell samples. The most widely spread method is to observe samples with a microscope. New methods allow us to scan a sample into a very large and detailed image. These images can be uploaded to the Cytomine web application and doctors can go through the images with ease. The application includes many functionalities, namely the ability to annotate regions of interest. Currently the ULiege MOOC server is used for the benefit of the students studying the medical field at the University of Liege. The students study particular images and are then evaluated at the end of the year. Meanwhile, the Cytomine app has been collecting data on the students' time spent on the website. The bulk of the data collected consists of where the students decided to look in the images (Gaze data). With that, attempts were made to find correlations between students' behavior and the results they obtain during exams. Using Machine Learning techniques, the goal is to predict a student's grade based on how they used the application. Currently, the model contains 395 students with over 2000 features. Random Forest and Extra Trees learning techniques have been applied to attempt to predict grades. Otherwise, another goal is to visualize these patterns. The idea would be to generate Heatmaps of the students' gaze data (Gazemap). These Gazemaps would be included in the app and users can be given access to this information. This could give teachers the ability to keep track of the students' work.

# 2 Acknowledgments

TODO

# 3 Introduction

## 3.1 Cytomine

Open-source rich Internet application for collaborative analysis of multi-gigapixel images. Cytomine can be described by three main properties :

- **Open Source** : The source code is available to the general public, in this case with Github. It has an Apache 2.0 License, which is very permissive for any third party. It is also accompanied by documentation that describes the different modules and how to use them.

- **Open Company** : It's a non-profit company that contribute and promote the project.
- **Open Research** : Cytomine employs researchers at the Montefiore Institute of the University of Liege. They develop Machine Learning algorithms, image informatics, and Big Data modules. Cytomine also collaborates with other researchers.

Cytomine was developed to ease the analysis of multi-gigapixel images. These images can take Gigabytes of disk space. For most computers, displaying such images at full resolution is impossible. With Cytomine, images are stored in a server. These images can be viewed using the web application with any modern browser. Cytomine handles everything locally so that performance is not an issue for the clients.

Cytomine also comes with a comprehensive and robust API (Application Programing Interface). This API allows clients to fetch and send data to and from a Cytomine Server. This is very useful for data analysis, it's well organized so that researchers can find what they are looking for. There are also tools included that eases certain aspects of handling big data. For examples image annotations allows the user to create zones of interest. These zones can share characteristics and using learning algorithms, researchers can learn new zones of interest.

## 3.2 Cytomine for education

Cytomine can be used as a tool of education for many subjects. Since it's open source and well documented, it is possible to fork it and develop new methods and modules. This can be done for many fields in computer sciences including Machine Learning, Vision, Big Data, or even front-end programing.

This is not limited to the field of Computer Sciences. Fields that require the use of large high definition images can benefit from using Cytomine. This includes astrology where, people can learn about planets and stars using imagery. Other sectors include Geology, Art, and in this case Medicine.

## 3.3 MOOC Server

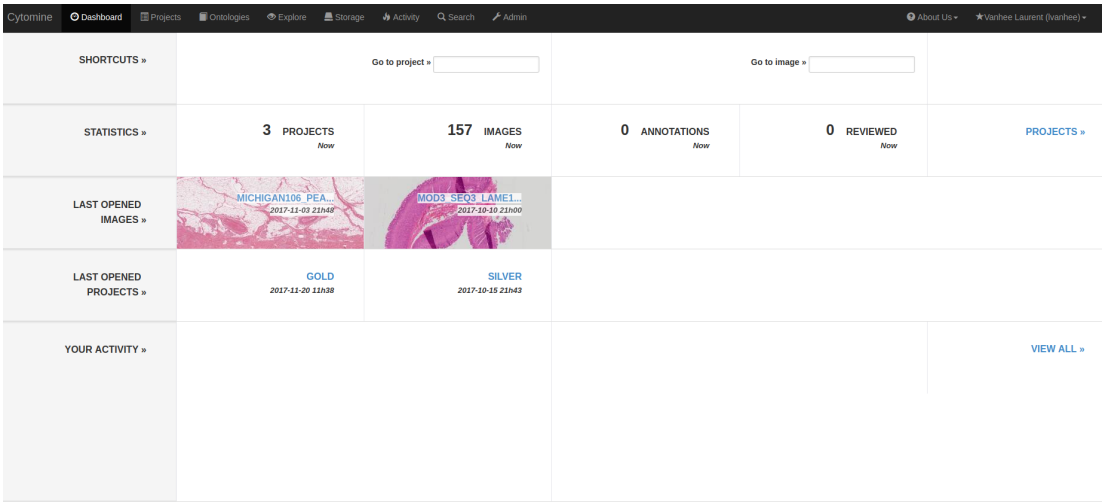



Figure 3.1: Cytomine MOOC home page

The MOOC server is a Cytomine Web Application used for education by the Faculty of Medicine at the University of Liege. It is used for the HISL0541 (General histology and alternative experimentation methods that do not use animals) course. This course is organized for students enrolled for a Bachelor in Medicine. Students have to study tissue samples. (Figure 2.2)

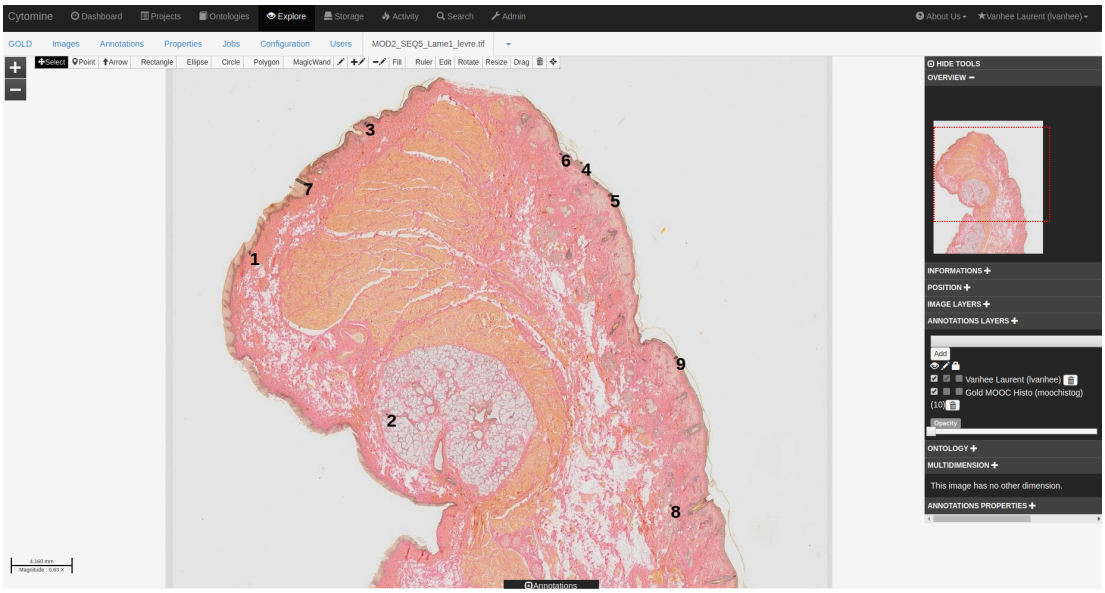



Figure 3.2: Cytomine MOOC Image example

Cytomine is a great tool for this particular field because it can be used to store and display high quality images of tissue samples. It is also easily accessible for the students, they are not restricted to only using it during lab sessions. It's available 24 hours a day,

therefore it gives students a good amount of liberty. The details of the assignments are given by Ecampus which is communication tool for students and teachers delivered by the University. The teachers can also use Cytomine to explain certain visual concepts in real-time. They can also use certain images for exercises by telling students to find patterns in those images.

Since the MOOC is completely integrated to the course, it directly impacts how students learn. Their exams are often based on what they learned using the tool. The goal is to see how much of an effect Cytomine has on students.

## 3.4 Teacher Input

The development of new tools and methods have heavily influenced the ways teachers approach their courses. This includes the sharing and consultation of resources from a distance. Every year, over 500 students from the Faculty of Medicine at the University of Liege partake in this new learning approach. The team working on the courses wanted to extend that knowledge and share it to a wider audience. This is where the MOOC server becomes interesting. MOOC stands for Massive Open Online Course. The courses that use the MOOC encourage democratization of the sharing of knowledge. The free access to high quality educational resources is now possible. In fact, there are platforms (for example : https//www.fun-mooc.fr) that share multiple MOOCs to encourage this principle. The MOOC is used in parallel with traditional courses. With the HISL0541 course, the teachers divided the contents into 6 modules. The first being a introduction to histology, and the rest a more concerns the vast family of tissue. They invite students to explore and observe multiple images. They offer two projects (GOLD and SILVER) that people can participate in (not only the students that follow the HISL0541 course). At the end, these participants are offered to partake in activities that can give them certificates. If the participants succeed with a score of 70%, they are given a certificates. Over 5000 participants that range from young to old, from students to employed workers, and from many different countries have participated in the MOOC. Overall most participants seemed very motivated and satisfied with this approach. This sharing of knowledge in a interactive and accessible way is great for anyone that have interests in this topic.

## 3.5 General Goals

The Main objective is to observe how the students use Cytomine. Since Cytomine is still in development, an objective is to make it robust and user friendly. This analysis can give some insight on what works well and what doesn't. As a follow up, it would be interesting to see the patterns that that can be observed in regards to the user. The teachers also provided some information on students that followed the course. This includes the grade obtained and all the tests that lead to that final grade. With this information, machine learning can be used to analyze these grades and what patterns lead to them. This analysis returns statistics in many forms since there is such a wide range of variables. (Figure 2.3)

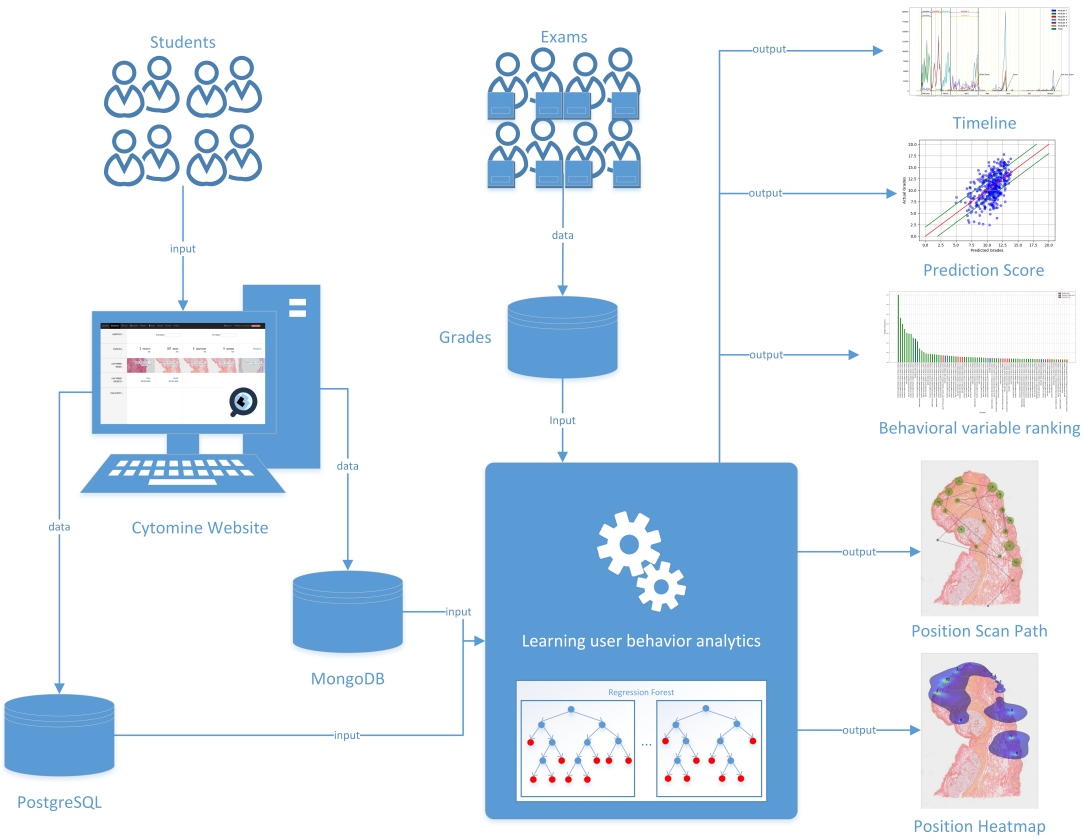



Figure 3.3: High Level View of the Tool

# 4 Tools and Methods

## 4.1 Cytomine Databases

TODO

## 4.2 Cytomine API

TODO

## 4.3 About ExtraTrees

TODO

## 4.4 Three Separate Components

Since there is a vast amount of data, most operations require some good amount of time to complete. Therefore, the data analysis tool is divided into three components :

- Data Acquisition.

- Data Manipulation.
- Data Learning.

### 4.4.1 Data Acquisition

During the 2016-2017 Academic year, Cytomine tracked and stored user information on the MOOC server. To obtain all this information, Cytomine is accessible by a REST API. To ease the access to the server, Cytomine developed a Python Client that was used for the acquisition of relevant information. With administrator rights, a user can get a hold of all if not most of the data stored on the SQL and MongoDB databases.

There are currently a total of two projects called GOLD and SILVER that students could participate in. To start of, each project contains a set of images and a set of students that have signed up.

| | Number Of students | Number Of Images |
|---|---|---|
| GOLD | 395 | 78 |
| SILVER | 85 | 75 |

Both projects have their own objectives but GOLD is more complete and thorough. unlike the SILVER project, the students were tested and graded on the course that was given to them but also the content of the project.

To analyze user behavior, it is necessary to fetch data that is relevant, this includes (as shown in Figure 3.3):

- **Resized images :**

  The images stored on Cytomine are in fact very large. For the analysis, a copy of the images will be useful. Many operations in the Data Manipulation component rely on the image resolution when it comes to complexity. It is important that the image is big enough to be viewable while small enough so that the operations done in a reasonable amount of time. The image downloaded is therefore rescaled to a maximum width and height of 1024 pixels.

- **Reference Annotations :**

  When observing images, students are usually given guidelines in forms of image annotations. These annotations are zones in the image that contain information that students can learn from. These annotations are given a number. Annotations from the same image have a different numbers. This represents the recommended order the user can traverse the image. In this study, only the geometrical center of the annotation is kept. In most cases, this loss of information should have no impact because the size of the zones are usually a couple pixels wide when put in the resized image. Reference annotations are drawn similarly to the example in Figure 3.1 that are indicated by numbers.

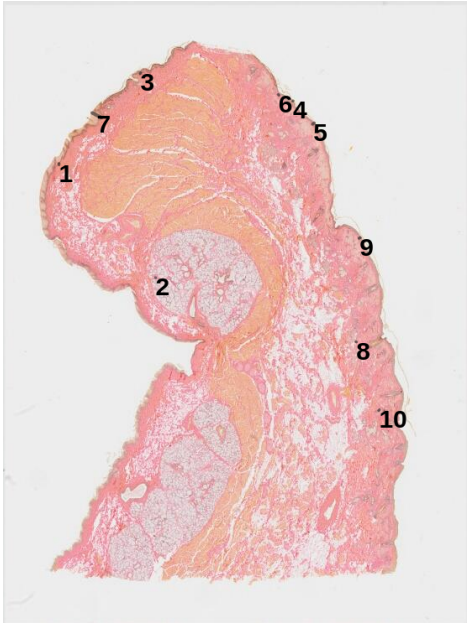



Figure 4.1: Example of Reference annotations

- **User Annotations :**

  Teachers can set annotations as guidelines, but normal users can also create annotations. If a student user notices something interesting on a patch of an image, that student can annotate it. Later, that student could for example approach a teacher with a question and use the annotation as a reference. Unfortunately, there are currently no User annotations. This will be discussed in section 5.

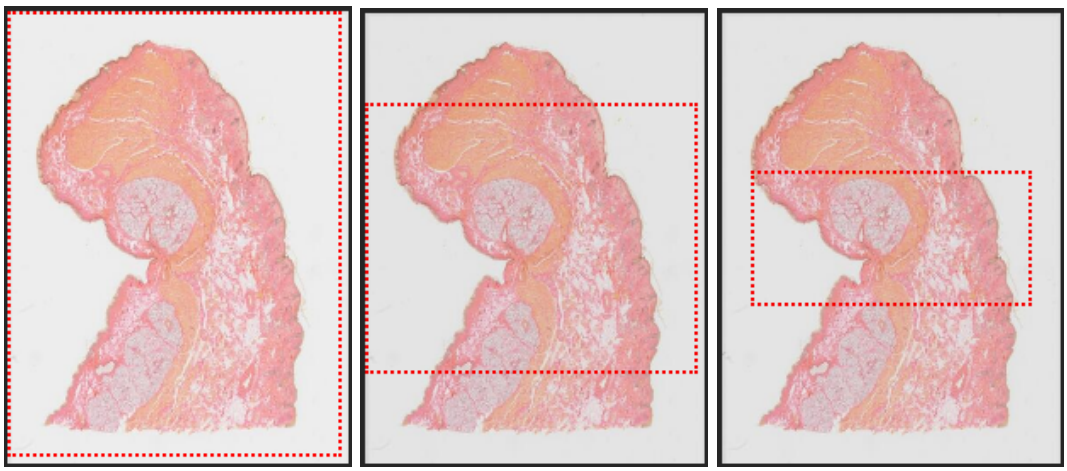
- **User Positions :**

  The most important information. A Positions is what the user sees at a current time stamp. Positions are defined by its center, four corners, time recorded, and zoom. Positions are saved on a regular basis when a user observes an image. More precisely positions are saved :

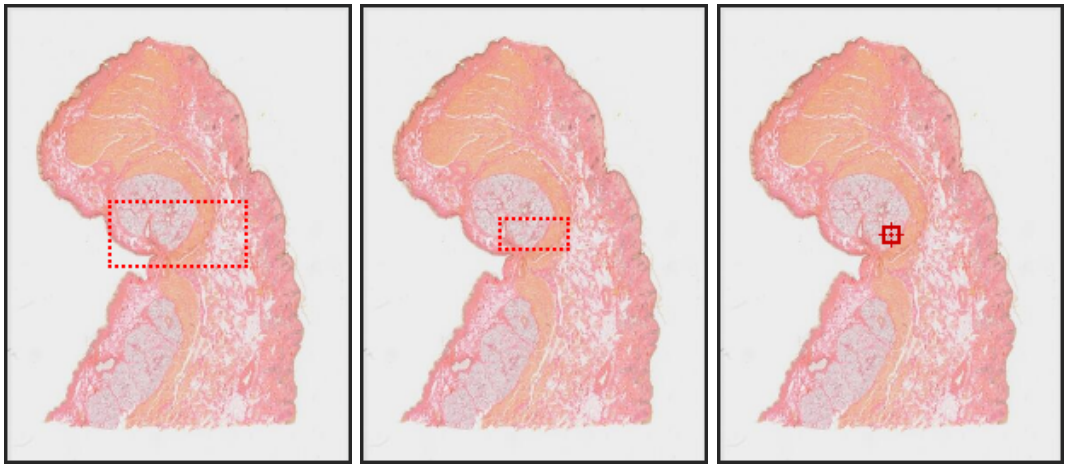
  - Every 5 seconds.
  - When the user switches zooms.
  - After the user finishes a movement on the image.

  Due to how frequently positions are recorded, this information comes in large quantity. Unfortunately, with all browsers (E.G. Google Chrome and Firefox) positions are still recorded when the page is minimized or tabbed out. This means that users can browse some unrelated website or even read course's Syllabus while positions are still recorded on Cytomine. The issue is that it's hard to determine whether or not a position is considered invalid because the user was not focused on it during the time. An idea

that was attempted was after a certain amount of positions with the same coordinates, is to remove the following positions with that coordinate. Unfortunately, it requires the setting of a threshold. This threshold would be arbitrary. After some attempts, this loss of information became a detriment to the Machine Learning and the results were worse. Therefore, the idea has been put on hold. A visualisation of the different zooms can be shown in Figure 3.2.



(a) Zooms 1 to 3 (b) Zoom 2 (c) Zoom 3



(d) Zoom 4 (e) Zoom 5 (f) Zooms 6 to 9

Figure 4.2: Figure of different zooms

- **Annotations Actions :**

Annotations are clickable. When clicked, a toolbox appears giving more information on the annotation. This action is also stored on the server. For the data recorded in 2017, an annotation action only contains a time stamp. It is only in later versions of Cytomine that the reference annotation identifiers were tracked with the annotation actions. In the case where the referenced annotation is unknown, it will be guessed

based on positions that appear at the same time.

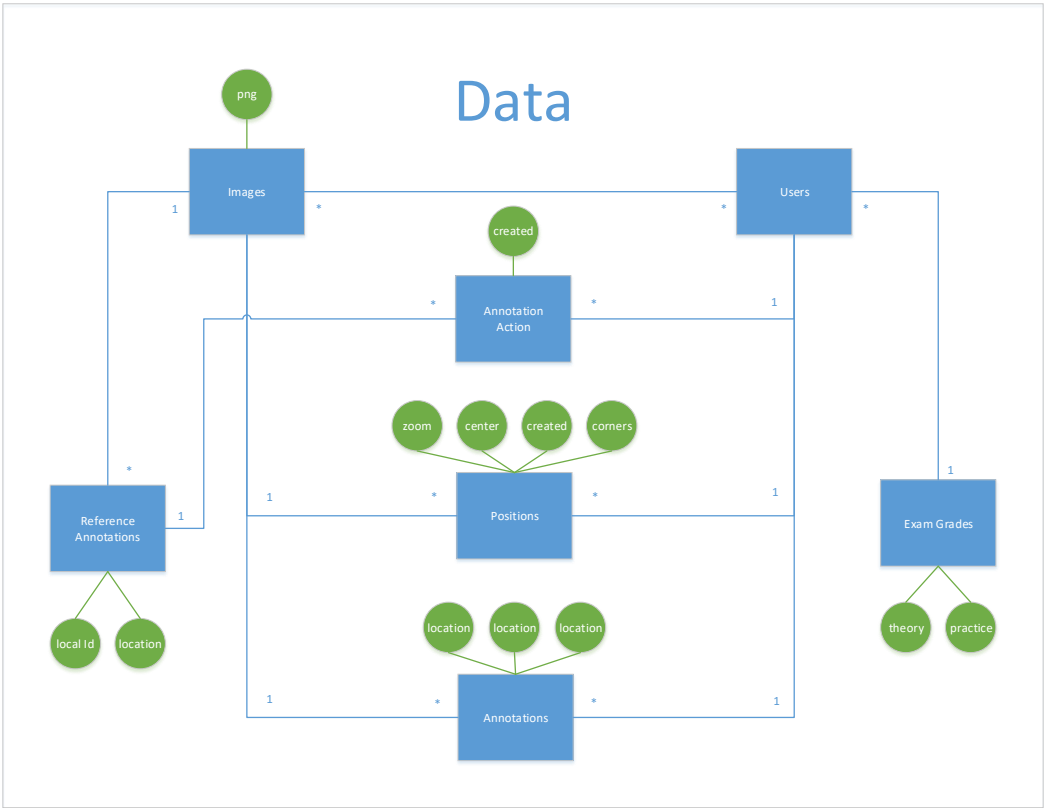



Figure 4.3: Data and their relations

All this information needs to be downloaded. Unfortunately, this can take up to 8 hours using the API while putting stress on the MOOC server. After some interruptions to the service, it was decided that the MOOC needed to be installed locally with backups of the original.

For both projects, an excel file containing user information not found on Cytomine was given. For the SILVER project, this only included basic information that were irrelevant to the analysis (names, emails, etc..). Meanwhile for the GOLD project, the University of Liege students had to partake in multiple graded assignments. The excel file given for the GOLD students therefore contained grades for numerous activities. These grades will be useful in order to understand some aspects of the user behavior on Cytomine.

The project has a dedicated directory to store all this information, usually in a csv format. For each data type, image, and user triplet, there is a dedicated file containing this information. These files will be opened and read by the Data Manipulation Component for generating statistics and ways to represent data visually. These files are organized in directories. This component, builds an underlying structure that will be used and manipulated by the other components. (Figure 3.4)

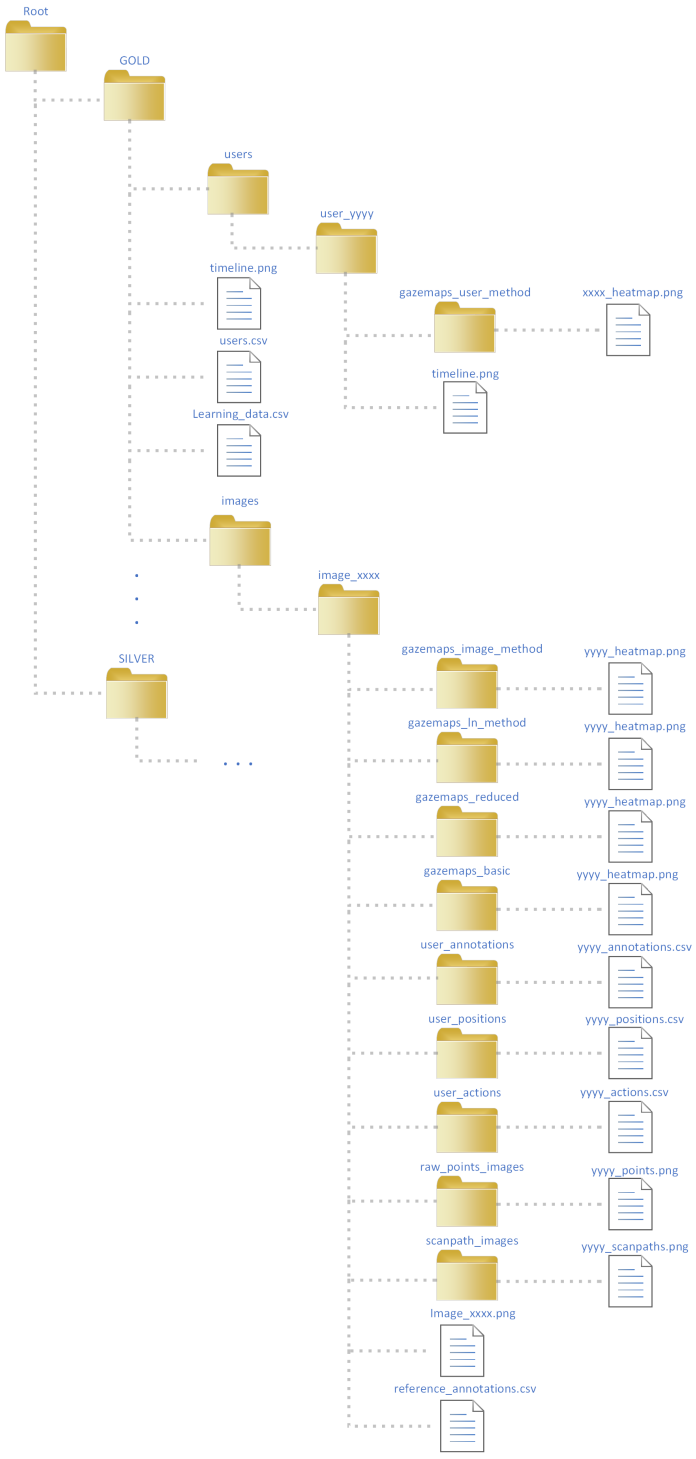



Figure 4.4: Directory and File Structure

### 4.4.2 Data Manipulation

This component is used to interpret the data that's in its most basic form. It has multiple outputs based on the parameters set. Since it's important to obtain information as a whole on the data set, the component is divided into two main parts:

- **Loading data to Memory:** The positions, the annotations, the actions, and the image information data are all loaded to memory.
- **Data Handling :** Calling numerous methods that extracts important information.

This component knows and uses the file directory structure. This allows it to find the files necessary to do the analysis. An object oriented approach was used in order to handle this information. In fact, there are multiple classes:

- **Data_Manager :** This class defines an object that will coordinate between the other classes in order to obtain the valuable information necessary.
- **Image_Data :** This class defines objects that will contain references to information on a specific image. It also comes with operations associated to said image.
- **User_Data :** This class defines objects that will contain references to information on a specific user. Similarly, to Image_Data, it comes with numerous operations.
- **Module_Data :** This class defines objects that will contain references to information on a specific module. Modules are predefined by the teachers as a set of objectives within a certain timeframe. There is a subset of images that are associated to a module. It therefore contains references to users and images and it also comes with its own operations.

These instances contain references to richer information. This information is handled using dictionaries for fast and easy access. For the set of positions associated to a user and image pair, the dictionary contains these key and value pairs:

- **'x' :** Array of X coordinates for all the positions.
- **'y' :** Array of Y coordinates for all the positions.
- **'dur':** Array of the duration for all the positions.
- **'timestamp':** Array of the exact time the positions have been recorded for each position.
- **'zoom' :** Array of the zoom values for each position (1 to 10)
- **'corners'** : Array containing the four corners for each position. The four corners are stored as an array with 4 values. These values contain a pair for the X and Y coordinates.

It is noted that the positions are sorted in regards to the timestamp value. This prevents many problems and makes specific operations much easier. The dictionary follows a similar structure for annotations that are either associated to an image and user pair (user annotation) or just an image (reference annotation):

- **'x' :** Array of X coordinates for all the annotations.
- **'y' :** Array of Y coordinates for all the annotations.
- **'id':** Array of the identifiers for all the annotations.
- **'localId':** Array of the local identifiers for all the annotations. Used for reference annotations. This property represents the annotation number shown on the images but also the recommended order of passage for these annotations
- **'type' :** Array of the type for all annotations. This is either a point or a polygon.

There are much fewer annotations then there are positions. Finally there are also annotation actions for image and user pairs:

- **'id' :** Array of the identifiers of the concerned annotations for all the annotation actions.
- **'action' :** Array of the action carried out for all annotation actions. The value is always 'select'.
- **'timestamp':** Array of the time the annotation action was carried out.

It is important to know that during the period of January 2017 to September 2017, annotation actions were collected but they did not specify which annotation the action was related to. To determine that the component guesses the annotation. The component finds the position with the closest timestamp and guesses the annotation closest to that position. It's not perfect but it guesses the most likely annotation since at most zooms the user can only see one annotation.

This data is referenced by the proper Image_Data and User_Data objects for easy access. When loading all this data into memory, extra operations have been carried out. The most notable being generating the gaussian surface for each zoom associated to each image. This uses a 2 Dimensional Gaussian distribution function. The Gaussian plane respects the dimensions of the positions associated to that zoom. This allows to better represent a position by not just taking into account the center and corners but the entirety of the field of view of the person. The goal is to give more importance to the center pixels of a position over the pixels near the edge. This will be very important when deriving statistics generating heatmaps.

When generating the gaussian with a height of $h$ and a width of $w$. The standard deviation set for the height and width are $s_h = \frac{h}{6}$ and $s_w = \frac{w}{6}$. $h_0 = \frac{h}{2}$ and $w_0 = \frac{w}{2}$ are the center coordinates of the surface. $coeff$ is a coefficient based on the zoom. The coefficient is a value from 0 to 1, the higher the zoom the higher the coefficient with $coeff = \frac{zoom}{max\_zoom}$. The equation is :

$$P(i,j) = \exp(\frac{-1}{2} * (\frac{(i-h_0)^2}{s_h^2} + \frac{(j-w_0)^2}{s_w^2})) * coeff$$

With $i \in [0,h]$ and $j \in [0,w]$.

With the standard deviation set, 99 percent of the gaussian distribution is in the surface. Therefore when calculating the center point, there is an output close to 1. Likewise, when calculating a point near the edge, there is an output close to 0. An interesting idea that was not implemented was to was to replace $coeff$ with $\frac{1}{2\pi * s_h * s_w}$. With that, the integral of all the Gaussian surfaces would be equal to 1. The problem with this method is with each zoom level, the height and the width halves in size. The surface would be four times as small. Therefore, from zoom 1 to 10 there is a factor of $4^{10}$ or about a million. The difference in values between zooms would be too extreme. This is why a more linear coefficient is used so that Heatmaps are observable. In the end, the idea is that the human attention is more focused on the center. Therefore a Gaussian distribution is ideal for this particular problem. In the following example is shown the Gaussian distribution for a surface with a length and width of 100, and a coefficient of 1. (Figure 3.5)

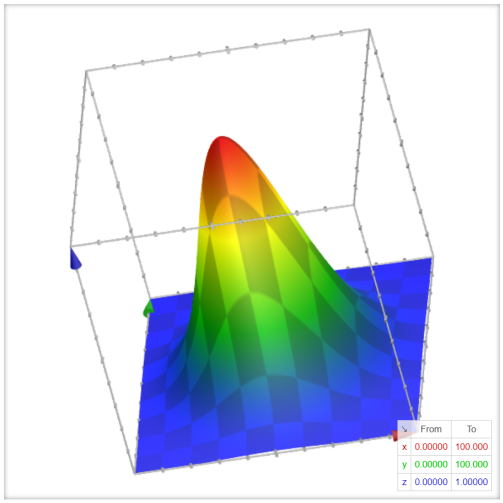



Figure 4.5: Gaussian Distribution Example

Of course, because the work is done on images, there is a limited resolution and the Gaussian is not as smooth. This is because each pixel is a point and a point has the value set by the formula.

Once all all this information is stored in memory. it is possible to call numerous tasks to analyze the data and draw statistics. This includes :

- **Output a Feature File :** Students are analyzed individually in regards to the images they visited. A set of features is then extracted in regards to their behavior. (see 4.2 for more information on the contents and structure of the file)

- **Output activity over time figures (Timeline) :** These figure represents the activity over time of a set of students or simply an individual in regards to different modules. It gives a good insight on habits and work consistency of the students. These figures depict the number of positions for every single day. The weekends are tinted in a light yellow. It also shows the activity in regards to the modules. Since the modules come with a set schedule, the figure also shows when they start and when they end. Finally, the dates for the exams are also a big focus. This allows the observation of the students' patterns as a whole or comparing the behavior of particular students. An interesting experiment would be to compare the figures for the very best student and the very worst students to get more insight. The example in Figure 3.8 shows the total activity for all the students.

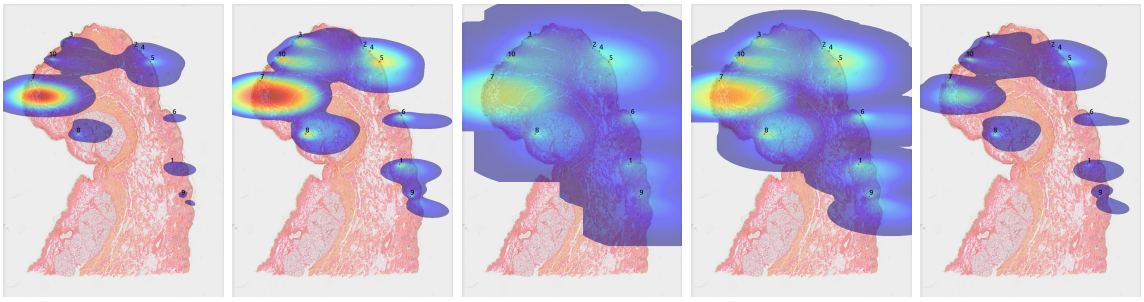
- **Output Gaze Map figures :** These figures represents the Gazemaps of the student activity in regards to an image. Since an image is resized to a reasonable resolution, each pixel is assigned a weight based on the user's activity near that pixel. Therefore, operations are done pixel by pixel. The values assigned rely on the 2D Gaussians generated for the positions. The 2D grid for the heatmap was generated with this component but generating and saving the image was implemented by PyGazeAnalyzer. Their heatmaps are beautiful but they lack information. For a pixel the color red means that it was the most visited a blue means that it was not a big focus. Even with a legend it would be hard to compare different images because the scales would not be the same for many heatmap implementations. Depending on the method, generating a heatmap can be a very costly operation. Methods include :

  - **Raw Gazemap :** For each position, the respective gaussian values are added into the heatmap at the right position. There are some advantages including that it's the fastest and shows clear distinctions between parts of the images. Drawbacks include that it is really hard to quantify and compare between users and other images. An another issue is with extreme cases the distinctions between regions of the image that been visited are too strong. For extreme cases, the most visited region might have such a high heatmap value that the other regions are hard to compare with each other.
  - **Logarithmic Gazemap :** The heatmap is generated similarly but a $Log_{10}$ normalizer applied at the end for each pixel. The advantages of that is that it's easier to compare areas of an image since the distinctions are smoothed. But it's main drawback is that it's really hard to quantify because the normalizer can be too extreme. Similarly, it's not straightforward when comparing students and images with each other since the scales are different for each heatmap.
  - **Logarithmic Gazemap Relative to an image :** The idea would be to reimplement the previous heatmap for every user of a specific image and standardize the scale. Since the minimum heat value will always be 0, this method only looks for the maximum heat value between all the heatmaps for that image. This value is set for all the heatmaps and they are drawn. This method's goal was to compare different students for the same image. Unfortunately, it still shares most of the same drawbacks of the Logarithmic Heatmap.
  - **Logarithmic Gazemap Relative to a user :** This is the same concept as

previously but the goal is to compare activity of the same student across multiple images. This has the same advantages and drawbacks as the previous heatmap.

– **Reduced Gazemap :** A interesting observation is that it's hard to quantify user observations for long durations. The longer a person looks at the same area the less important the positions become over time. This explains the attempt to normalize the heatmap using the logarithmic function. The problem is that there's no theoretical explanation to why the the logarithmic method should work. In this case a pixel is represented by vector instead of a value. The vector contains a list of all the respective Gaussian value for each position that is close enough. The value is always in the range of $[0, 1]$. For each pixel, the list is sorted inversely. Finally, the value of the pixel is given by a weighted sum of the values of this list. The weights follow a geometrical sequence that converges. Therefore, after numerous positions, the weights are near 0. This normalization method is further explained in the section 4.2.3. In this case, $w$ is set to 0.95 and the sequence converges at a value of 20. Therefore, the heat values from each pixel belong in the range $[0, 20]$. The maximum value for all heatmaps can easily be set to 20. This method has a more relaxed normalization and it is much easier to compare different heatmaps and infer heat values by looking at the image. The main drawback is that it's much slower. Since for each pixel, the algorithm is given a list to sort a values to weight it takes much more time.

Figure 3.6 compares the Gazemaps of different methods for the same user and image pair. In this example, the student focused a zone with a low zoom for a very long duration while also visiting annotations for a reasonable amount of time. Observing these Gazemaps, they all have their flaws. The logarithmic normalizer is too powerful. Even though the colors for the reduced Gazemap is lighter, the colors are consistent with every other reduced heatmap that can be generated. Unlike the two relative heatmaps who are only consistent between either users in the same image or images associated with the same user.



(a) Raw (b) Logarithmic (c) Image Relative (d) User Relative (e) Reduced

Figure 4.6: Example Figure of different Gazemaps

- **Output Scan Path Figures :** These figures represents the path taken by the students when they scan through an image. Another Idea that seems interesting is to observe a student's viewing order when looking at images. The idea is to draw arrow an plot that connects each position to the following position. After doing so for all the positions,

it's possible to analyze the viewing order. Unfortunately, in most cases there are too many positions. This is why clustering is used to group up close positions in regards to time and space. K-Means clustering was implemented with K predetermined and ranging from 10 to 50 (K = 20 ideal). With this amount of clusters there may be a small loss in information but it allows the Scan Path to be observable. These clusters are arranged by the average time of their positions. Each cluster is also given a weight based on the number of positions that belong to it. This makes the scan path complete in a sense that it shows most of the needed information relating to the user's viewing order. The example in Figure 3.7 shows the scan path associated to the same image and user pair as the Figure 3.6. This shows that the zone where the student focused for a long duration was during the begining of the scan. The student may have been reading up on the course while keeping the image still. Following that, the student started visiting the different annotations. It's interesting to see that the student did not respect the order for the most part. The scan order is annotations 7, 8, 3, 2, 4, 1, 6, 5, 10,and finally 9.

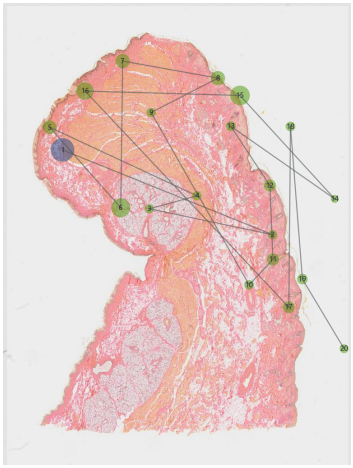


Figure 4.7: Scan Path Example

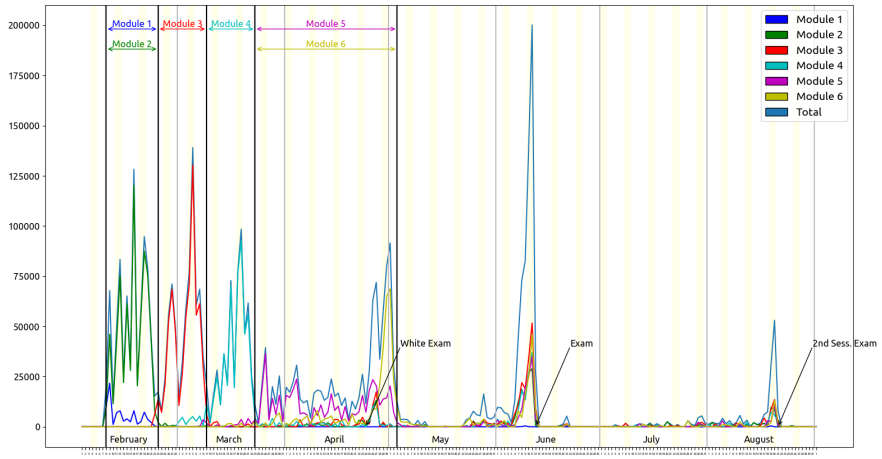



Figure 4.8: Overall Student Activity over Time

### 4.4.3 Data Learning

This component has the objective of working with the output Feature file from the Data Manipulation component. As there many output variables, a Graphical Unit Interface (GUI) was implemented. This GUI shows the list of output variables that the user can chose from. The user then has the option to run two different methods :

- **Building a full model :** All the feature variables have been fit into a ExtraTree Regressor with the selected output variable. From this model, statistics are extracted and visualized. This includes :
  - **Feature Importance Histogram :** An ExtraTree Regressor outputs the impact of all the variables after the model is fit. Since there are over two thousand features, they can't all be visualized. Therefore, only the 80 most impactful features have visualized with an Histogram. Features are divided into 3 categories, those who were associated to specific Modules (in red), those who were associated to specific images (in green) and the more general features (in blue). This helps give a rough idea on the effect of certain features.
  - **Image Importance Histogram :** This Histogram uses the features associated to specific images. For each image, the average feature importance is calculated. This is then shown in the Histogram. Images were also split into categories based on the module they belong to. The each category is represented by a color. This helps determining what images and modules had the biggest impact on the grades obtained.
  - **Variable Correlation Scatter Plot :** This scatter plot compares the output variable to a top feature. This visualizes the relation between those two variables. This includes aline with a Pearson Correlation to represent the slope of the of the relation.

- **Running a Leave-One-Out Cross Validation :** With a total of N individuals, the model is fit N times with N-1 individuals. For each iteration, the individual left out will be tested with the model. With this, the learned value is returned and is compared with the original value. After each iteration, statistics were derived from the results :
  - **Mean Absolute Error (score):** The average absolute error obtained by the cross-validation.
  - **Score of the Median Model :** The cross-validation is compared to a method that does a leave-one-out cross validation of the set by just taking the median value. This is to give confirmation on whether or not the results of the ExtraTree Regressor are determined somewhat randomly.
  - **Score Difference :** This is the difference between the two previously mentioned scores. The higher the difference, the better.
  - **P-Value :** The P-value of the cross validation results with the original output variables. This method uses the T-Test to see if the two variables are significantly different from each other. The P-value is valued from 0 to 1 with 1 meaning that the sets are exactly the same and 0 meaning the sets are completely different. In this case, the ideal P-value would be 0.95.
  - **P-Value of the Median Model :** Similar to the previous P-Value but applied to the median model.

  This leads to the implementation of statistics that are visualized in graphs:

  - **Scatter Plot :** This compares the learned results with the output variables. It includes four Lines. The first line (in red) represents $x = y$, the best possible outcome for the learning. The goal is to have the most accurate regression. This happens when the scattered points are closest to that line. The next two (in green) represents $x = y \pm score$, this is to give a interval where the error for the scattered points inside are less than the mean. The last (in blue) represents the slope of the relation between the two variables. It includes a Pearson Correlation to better understand the slope of the line.
  - **Box Plot :** This visualizes the error between the learned results with the output variables. In this case, it draws the median error, the two quartiles Q1 (seperating the 25 % with the lowest error with the rest) and Q3 (seperating the 75 % with the lowest error with the rest), and the two whiskers. The whiskers have a reach that are 1.5 times the inter-quartile range. This means that any error outside the reach of the whiskers would be deemed extreme cases. These cases are related to really high error values.

The ExtraTree (Extremely Randomized Tree) regressor is a class that fits a number of randomized decision trees on many sub samples of the population. This method uses averaging to to improve learning accuracy all while controlling over-fitting. It is therefore considered a Ensemble Method. The Regressor has been implemented with sklearn (http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesRegressor.html). It has been configured with the following parameters :

- **Number of Estimators : 10 000**
  Represents the number randomized trees generated.

- **Max depth : None**
  The max depth of the trees.
- **Bootstrap : False**
  Whether or not bootstrapping was implemented. Bootstrapping implements sampling with replacement within the training set.
- **Max Features : Auto ( = Number of Features)**
  Max number of features that can be used by the randomized tree.
- **Max leaf Nodes : None**
  The maximum number of leaf nodes that be present in a tree.
- **Minimum Impurity Decrease : 0.0**
  A node will be split if the impurity is less than the minimum impurity decrease value. Threshold for early stopping tree growth.
- **Minimum Samples Leaf : 1**
  The minimum number of samples required at a leaf node.
- **Minimum Samples split : 2**
  The minimum number of samples required to split an internal node.
- **Minimum Weight Fraction Leaf : 0.0**
  Minimum weighted fraction of the sum total of all weights (of all input samples) required at a leaf node. If 0.0, samples have equal weights.
- **Criterion : msq - Mean Square Error (mae is too slow)**
  The method to measure the quality of a split.
- **Out-of-bag Score : False**
  Whether out-of-baf samples are used to estimate how well unseen data is fitted to a regression line.

Since the ExtraTree regressor handles overfilling by itself, it's useful to have many estimators when handling a vast amount of features. Also, with many estimators the method will converge to output the best possible results where ExtraTrees are are concerned. It is one of the more ideal choices for such a data set for many reasons :

- The high number of features paired with a small number of individuals. Ensemble methods are ideal in this scenario.
- Easy to interpret (variable importance for example).
- The data comes all at once, so the tree only needs to be fit once.
- Good for modeling linear interactions. Most relations studied in this project are linear. With the sample size, it's risky to try and infer more complex relations.

# 5 Data Analysis

## 5.1 Experiments

Even though the data was obtained for both the GOLD and SILVER projects, the experiments were on run on the GOLD project. This is due to the many constraints given by the SILVER project including the sample size and the lack of results (student grades). The goal for most experiments is to learn a specific grade the student obtained based on all the information gathered in the Data Manipulation component. There were over 2000 features generated, where each one can weigh in on the prediction. These features were used to learn over 13 grades including the final grade obtained by these students (first session).

Due to the nature of the dataset, regression trees were the most ideal. The somewhat small dataset paired with a large amount of features is a big constraint to work with. Therefore, Ensemble methods were used and tested. This includes Random Forest regression and Extra Trees regression using sklearn.

The Learning component is given a statistics file containing rows of students paired with their features. The features are split into three categories :

- **M** : Meta-data variables, these variables have no statistical significance. These variables usually contain basic information on the users.

- **X** : Variables with statistical significance, these variables mostly include data extracted from the Cytomine website. Either individual image variables or variables on the set of images. These variables are used in the machine learning model as input variables.

- **Y** : Result variables, these variables is what the algorithm is attempting to guess using the **X** variables. These variables are used in the machine learning model as output variables. In this case, they are the students' grades.

With this, a couple experiments have been ran. For each Y variable (that will be described in the next sections), the model was built fully for the first analysis. The second analysis, a leave-one-out cross validation was ran. The goal is to give insight on different user behaviors. Hopefully, these experiments will prove to be useful.

## 5.2 Data Set

### 5.2.1 Students

Like mentioned earlier, there are a total of 395 students in the data set. Students are defined by their features and the grades obtained. These student followed the course HISL054, "General histology and alternative experimentation methods that do not use animals" at the University of Liege during the academic year 2016-2017. Most of the work done by the students is done online using the MOOC and Ecampus. Ecampus is a website used by students

and teachers of the University of Liege to exchange information used for courses. In fact all of the assignments for this course are explained on Ecampus. The issue with this is that it's a different entity from the MOOC and therefore is not ideal for fetching information. But it's not much of a problem because there's nothing to retrieve about students on Ecampus that is not already known.

When applying machine learning techniques, students with a 0/20 were taken out of the sample. This is due to the fact those students defaulted to signing the exam. Since there are too few that do so, there's not enough information to find correlations between people who sign and the features. Keeping students who sign also increases the error rate because they are extreme cases. The goal is to guess how well the exam would have went for the students. If the student decides not to bother with taking the exam, it's a lost sample.

### 5.2.2 Teacher Input : Grades

The students are evaluated by multiple exams and quizzes, these take the form of:

- **QCM : Multiple choice question test.** Students are given questions and have to respond with 1 out of the 9 (or less) options. They have to set a degree of certainty for each response. The exam results are then calculated by a machine. The students are also given more boxes to answer in case they make a mistake the first time.(Figure 4.1)


Figure 5.1: QCM sheet and question example

- **QCL : Identification and Incidence test.** Graded similarly to a QCM, the students are given multiple questions to answer and they have to fill out a form. The are a couple differences. Each question is split into two, the identification and the incidence. For the identification, the students have to identify an object on an image. They are given an exhaustive list of possible answers ranging from cells to tissues. Each answer contains is associated to a 3 digit code that they need to write on the form. For the incidence, the students needs to identify how the observed object "was cut". There are a total of 3 possible answers, transversal, longitudinal, and undetermined. The answer is written on the form under the identification answer.(Figure 4.2)

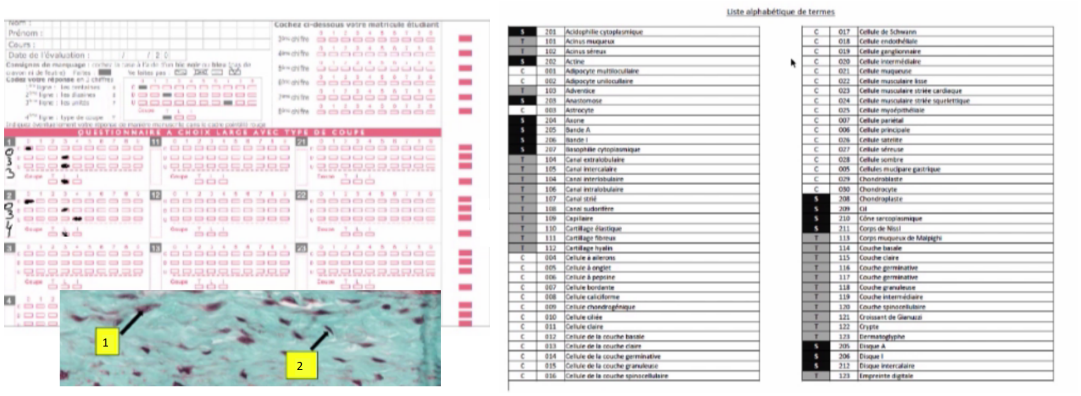


Figure 5.2: QCL sheet with question example and list of answers

- **QROL : Long answer open question test.** Students are to write and explain their answers in a detailed fashion. (Figure 4.3)

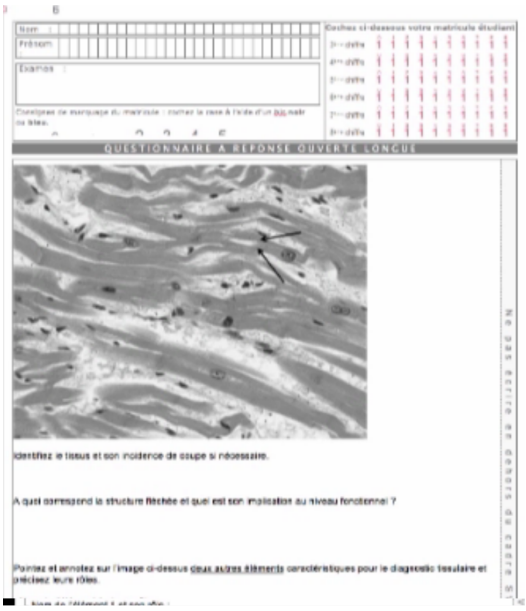


Figure 5.3: QROL sheet with a set of questions

These variables will be denoted as **Y** variables for output. Out of the 13 results associated to the students, 3 were white tests given as a practice tool :

- QCL identification white test.
- QCL incidence white test.
- Practical QCM white test.

Similarly, 7 graded exams and quizzes were given to students with 3 being theoretical and 4 being practical. Something to note for the practical exam is that there are 2 different exam

forms for the QCM and the QRL. This means that half the students are given different questions from the other half. The list of exams include:

- QROL1 theory (10%).
- QROL2 theory (10%).
- QCM theory (30%).
- QCM practical (20%).
- QROL practical (10%).
- QCL identification Practical (16%).
- QCL incidence Practical (4%).

Finally, based on the previous results given, there are:

- Total Theory (50%)
- Total Practical (50%)
- global Grade (100%)

Most of the experiments are done using one of these grades as Y variables. In later experiments, some variables will be set as features. The learning of final grades using white test results as bonus features (**X** variable) could yield better results. Learning theoretical grades while also using practical grades and vice versa can also give some interesting results.

Statistics can be derived for each variable :

| | White Test QCL ID | White Test QCL IC | White Test QCM | QROL Theory 1 | QROL Theory 2 | QCM Theory | QCM Practical | QROL Practical | QCL ID Practical | QCM IC Practical | Total Theory | Total Practical | Global Grade |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Minimum | 1.33 | 1.33 | 0.36 | 0.44 | 0.67 | 3.43 | 3.16 | 0.29 | 1 | 2 | 2.59 | 2.01 | 1.68 |
| Maximum | 18.67 | 20 | 19.64 | 20 | 20 | 17 | 17.32 | 18.82 | 20 | 20 | 17.8 | 17.99 | 17.85 |
| Average | 8.09 | 12.17 | 10.18 | 10.67 | 12.16 | 10.05 | 10.88 | 8.6 | 11.15 | 15.03 | 10.56 | 10.81 | 10.63 |
| Median | 8 | 13.33 | 10.31 | 10.75 | 12.83 | 10.03 | 10.8 | 7.93 | 11 | 15 | 10.75 | 10.76 | 10.67 |
| Variance | 14.09 | 17.67 | 13.3 | 16.78 | 19.51 | 6.41 | 7.93 | 18.92 | 16.48 | 9.35 | 8.46 | 9.92 | 8.64 |
| Standard Deviation | 3.75 | 4.2 | 3.36 | 4.09 | 4.41 | 2.53 | 2.81 | 4.34 | 4.06 | 3.05 | 2.9 | 3.15 | 2.94 |

Teachers also input basic student information, but it mostly consists of general information that won't be used in the experiments. These will be denoted as **M** variables:

- **ID Cytomine :** The Cytomine ID associated to the user. (Compulsory)
- **LAST NAME :** User's last name.

- **FIRST NAME :** User's First name (forename).
- **GROUP :** group user belongs in (GOLDULiege, GOLD, or SILVER). GOLDULIEGE is a subset of GOLD containing the set of students following the course. This will be used for the analysis.
- **USERNAME CYTOMINE :** user's Cytomine username.

### 5.2.3 Features pre-Calculated

There are over two thousand features calculated and generated by the Data Manipulation component. These variables belong to the **X** category and are listed:

- **NB IMAGES VISITED :** The total number of different images that a user has opened over the course of the year.
- **TOTAL NB POSITIONS :** The total number of positions obtained from all the images opened.
- **AVG NB POSITIONS :** The mean number of positions obtained relative to all the images opened.
- **MEDIAN NB POSITIONS :** The median number of positions obtained relative to all the images opened.
- **TOTAL IMAGE VIEWING TIME (s) :** Total amount of time spent viewing images.
- **AVG IMAGE VIEWING TIME (s) :** Mean amount of time spent viewing images.
- **MEDIAN IMAGE VIEWING TIME (s) :** Median amount of time spent viewing images.
- **NB POSITIONS AT ZOOM <x> :** with <x> between 1 and 10, represents the zoom level of a position. 1 variable per zoom value. It represents the total number of positions at zoom <x>.
- **AVG ZOOM :** The mean zoom level over all the positions collected.
- **MEDIAN ZOOM :** The median zoom level over all the positions collected.

- **TOTAL NB ANNOTATION ACTIONS :** The total number of times the user clicked on a reference annotation.

- **AVG NB ANNOTATION ACTIONS :** The mean number of times the user clicked on a reference annotation.

- **MEDIAN NB ANNOTATION ACTIONS :** The median number of times the user clicked on a reference annotation.

- **AVG NB POSITIONS AT ZOOM <x> :** with <x> between 1 and 10, represents the zoom level of a position. 1 variable per zoom value. It represents the mean number of positions at zoom <x> relative to all images visited.

- **MEDIAN POSITIONS AT ZOOM <x> :** with <x> between 1 and 10, represents the zoom level of a position. 1 variable per zoom value. It represents the median number of positions at zoom <x> relative to all images visited.

- **SCORE OF ANNOTATION <y> AT IMAGE <x> :** with <y> being the annotation identifier and <x> being the image identifier. When images have annotations, students tend to focus on these points. For each position, the program generates a 2 Dimensional Gaussian function to represent the position. The size and values of this Gaussian relies on the zoom of the position. For example at the center of the Gaussian with the highest zoom, the value is 1. While at the lowest zoom the value is $1/MAX\_ZOOM$.

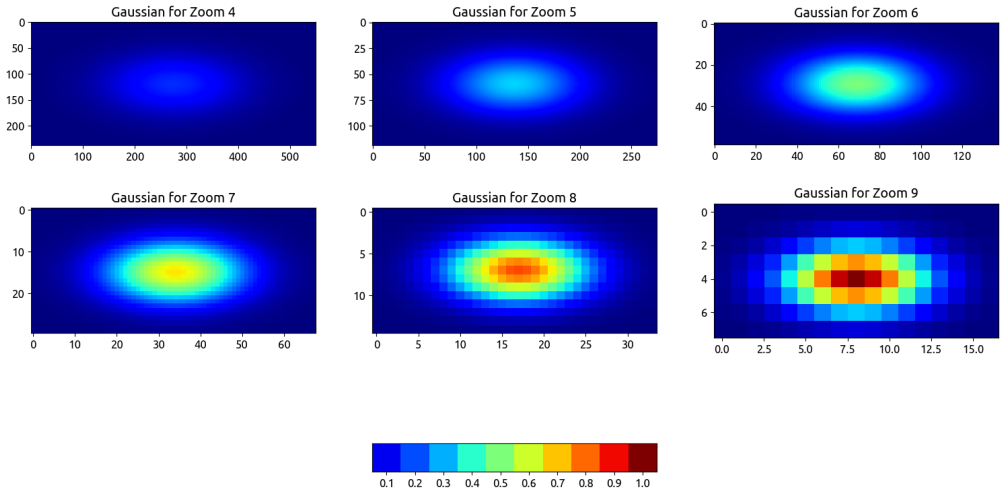



Figure 5.4: Example Gaussian Grids for an Image that has a zoom up to 9

So each annotation has a center coordinate. For this coordinate, a list of Gaussian values is generated based on all the positions near the annotation. This vector is sorted inversely. It's important to note that it does not take much time for the student to assimilate all the information on a part of an image. So the idea is to weight these positions to the point that after enough positions, the next ones would have a weight of close to 0. To do so, a geometrical sequence was used:

$$\sum_{i=0}^{N} w^i$$

N is the number of positions. In our case, $w$ was set as 0.95. This means that this sequence converges to 20, and after about 50 positions the sequence is close to the convergence value (about 18). Since the list of Gaussian values are sorted, the highest values will have the highest weight. The equation becomes :

$$score = \sum_{i=0}^{N} L[i] * w^i$$

With L being the list of Gaussian values for the annotation. This means that the highest value has a weight of 1, the second a weight of 0.95, the third $0.95 * 0.95$, and so on. The value calculated is then used for the score. If a student actually observed in detail the annotation, they would usually get values above 10.

- **USER SCORE AT IMAGE <x> :** With <x> the image identifier on Cytomine. The users are given scores between 0 and 1 (not opening an image gives the user a score of 0). The score represents how well the user observed the annotations in the given image. In short, if a user spends a good amount of time on top of an annotation or if the user clicks on the annotation, he/she will get points for that annotation. Doing so for all annotations gives a final score. The score for an annotation is given by the previous variable (SCORE OF ANNOTATION <y> AT IMAGE <x>) and re dimensioned so this variable returns a score between 0 and 1. If the image does not have any annotation, the scores for are calculated for the entire image and the average is returned.

- **AVERAGE USER SCORE :** The average of all the scores defined previously for a user.

- **NB POSITIONS AT IMAGE <x> :** with <x> being the image Identifier, represents the number of positions recorded at that image for a user.

- **TIME SPENT AT IMAGE <x> :** With <x> being the image identifier, represents the total time spent on an image for a user.

- **NB OF ANNOTATION ACTIONS AT IMAGE <x> :** With <x> being the image identifier, represents the number of annotation Actions at that image for a user.

- **NB OF POSITIONS WITH ZOOM <y> AT IMAGE <x> :** With <x> being the image identifier and <y> being the zoom value [1-10]. It Represents the number of positions for a certain zoom at that image for a user.

- **NB IMAGES VISITED DURING MODULE <x> :** The total number of different images that a user has opened that are associated to the module <x>.

- **AVG NB POSITIONS DURING MODULE <x> :** The mean number of positions obtained relative to all the images opened that are associated to the module <x> during the corresponding time period.

- **MEDIAN NB POSITIONS DURING MODULE <x> :** The median number of positions obtained relative to all the images opened that are associated to the module <x> during the corresponding time period.

- **TOTAL NB POSITIONS DURING MODULE <x> :** The total number of positions obtained from all the images associated to the module <x> during the corresponding time period.

- **TOTAL TIME SPENT DURING MODULE <x> (s) :** Total amount of time spent viewing images associated to the module <x> during its given time period.

- **AVG TIME SPENT DURING MODULE <x> (s) :** Mean amount of time spent viewing images associated to the module <x> during its given time period.

- **MEDIAN TIME SPENT DURING MODULE <x> (s) :** median amount of time spent viewing images associated to the module <x> during its given time period.

- **NB POSITIONS DURING MODULE <y> FOR IMAGE <x> :** with <x> being the image Identifier of an image associated to the module <y>. This represents the number of positions recorded at that image for a user during the module's time period.

- **TIME SPENT DURING MODULE <y> FOR IMAGE <x> :** with <x> being the image Identifier of an image associated to the module <y>. This represents the time spent at that image for a user during the module's time period.

- **NB ANNOTATION ACTIONS DURING MODULE <y> FOR IMAGE <x> :** with <x> being the image Identifier of an image associated to the module <y>. This represents the number of annotation actions recorded at that image for a user during the module's time period.

- **NB POSITIONS WITH ZOOM <z> DURING MODULE <y> AT IMAGE <x> :** with <x> being the image Identifier of an image associated to the module <y>. This represents the number of positions recorded at that image for a user during the module's time period with zoom <z>. The zoom value <z> ranges from 1 to 10.

- **AVERAGE ZOOM DURING MODULE <y> FOR IMAGE <x> :** with <x> being the image Identifier of an image associated to the module <y>. This represents the average zoom level of all the positions recorded at that image for a user during the module's time period.

- **MEDIAN ZOOM DURING MODULE <y> FOR IMAGE <x> :** with <x> being the image Identifier of an image associated to the module <y>. This represents the median zoom level of all the positions recorded at that image for a user during the module's time period.

- **NB POSITIONS AT ZOOM <z> DURING MODULE <x> :** with <z> between 1 and 10, calculated for all the images associated to the module <x>. This

represents the number of positions for a specific zoom for a user during the module's time period.

- **AVERAGE NB POSITIONS AT ZOOM <z> DURING MODULE <x> :** with <z> between 1 and 10, calculated for all the images associated to the module <x>. This represents the mean number of positions for a specific zoom for a user during the module's time period.

- **MEDIAN NB POSITIONS AT ZOOM <z> DURING MODULE <x> :** with <z> between 1 and 10, calculated for all the images associated to the module <x>. This represents the median number of positions for a specific zoom for a user during the module's time period.

- **TOTAL NB ANNOTATION ACTIONS DURING MODULE <x> :** With <x> being the module identifier, it represents the total number of annotation actions for a user during the module's time period.

- **AVG NB ANNOTATION ACTIONS DURING MODULE <x> :** With <x> being the module identifier, it represents the mean number of annotation actions for a user during the module's time period.

- **MEDIAN NB ANNOTATION ACTIONS DURING MODULE <x> :** With <x> being the module identifier, it represents the median number of annotation actions for a user during the module's time period.

- **AVERAGE USER SCORE DURING MODULE <x> :** With <x> being the module identifier, it represents the average predefined score for a user for the module's images during the respective time period.

- **USER SCORE AT IMAGE <y> DURING MODULE <x> :** With <x> being the module identifier and <y> the image identifier, it represents the predefined score at the image for a user during the module's time period.

- **SCORE OF ANNOTATION <z> AT IMAGE <y> DURING MODULE <x> :** With <x> being the module identifier and <y> the image identifier, it represents the predefined score of the annotation <z> at the image for a user during the module's time period.

- **PERCENT TIME WORKED AT NIGHT :** Value from 0 to 1. This represents the ratio of the time spent working from 6pm to 6am.

- **PERCENT TIME WORKED LATE :** Value from 0 to 1. This represents the ratio of the time spent working from 1am to 6am.
- **PERCENT TIME WORKED MORNING :** Value from 0 to 1. This represents the ratio of the time spent working from 6am to 12pm.
- **NUMBER OF DAYS WORKED :** The total number of days where a user opened at least one image.
- **PERCENT TIME WORKED DURING MODULE <x> :** When working on Cytomine, the user can open images associated during a module during its time period or outside of it. This represents the ratio of the user activity during the time period.
- **ANNOTATION <y> VISITED BEFORE ANNOTATION <y + 1> AT IMAGE <X> :** This binary variable determines whether or not a student visited the annotation <y> before the annotation <y + 1> for an image <x>. The users are encouraged to study the reference annotation in a specific order.

All these features end up describing how users participated on Cytomine and their effects will be analyzed. It will be evident that most of these features will have little impact in the learning process.

## 5.3 Results

### 5.3.1 Comparing Timelines

In this small experiment, there were a total of 8 timelines geneated. Four of those belonging to the students with the highest grades and four belonging to the students with the lowest grades. The simple objective is to compare the results from the two sets and also compare results within a set. (Figures 4.5 and 4.6)

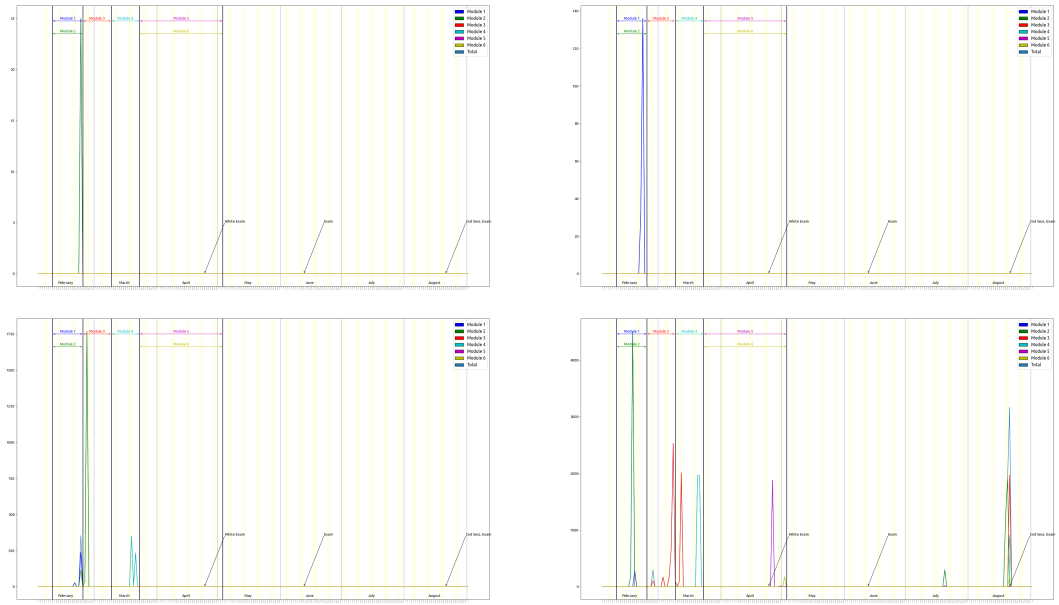


Figure 5.5: Timelines of the students with the lowest grades

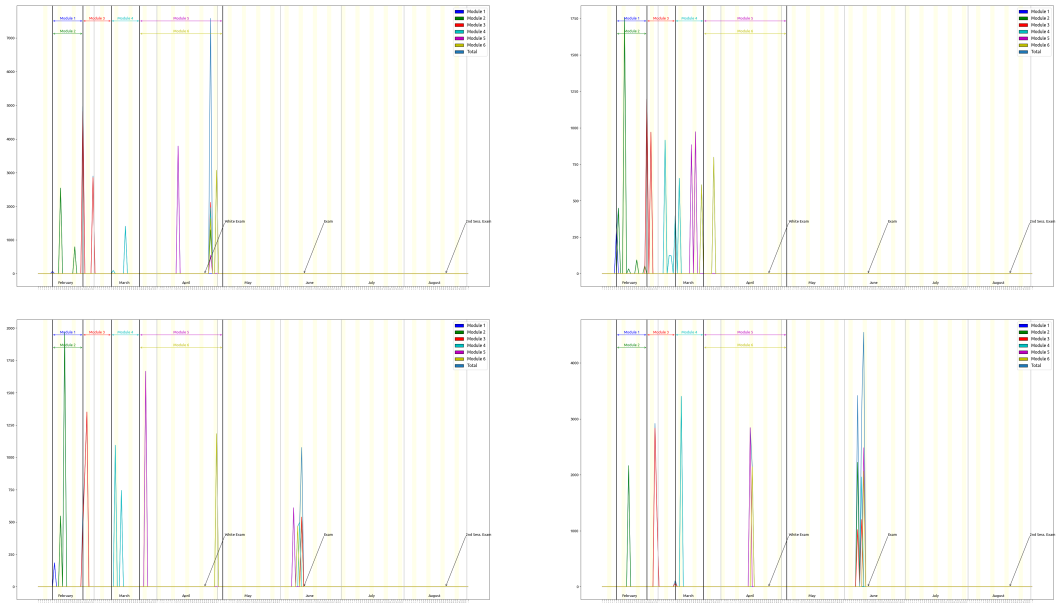


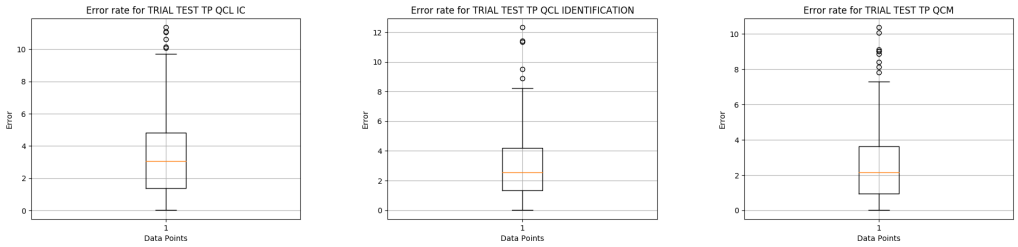
Figure 5.6: Timelines of the students with the highest grades

For the timelines of students with the lowest grades, for the mostpart it's pretty straightforward. There are are a small number of positions and some modules were not even visited. There is one student who followed different patterns with a respectable activity. Unfortunately, the student still failed the exam.

Meanwhile, for the students with th highest grades, there are many more positions as a whole. Also, images for all the modules were visited. There is a clear difference between the two sets. But in the set of students with the highest grades, there are many different patterns. Some students have a significant higher activity than others. One worked on modules ahead of time, and some didn't even log onto Cytomine the weeks before the exam. This shows that it's hard to judge a student's performance and comprehension of the course. Students follow different habbits and exhibit different patterns. There is no "best" method for studying according to this small study. The Machine Learning can give more insight on this question.

### 5.3.2 White Test Grades

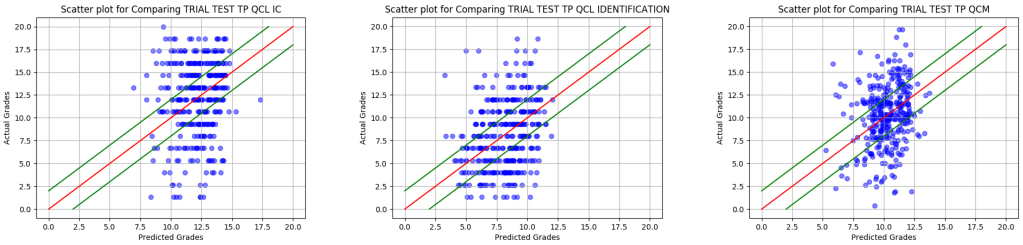
These 3 ungraded exam were taken halfway during the semester (24th of April 2017). They reflect what a student has learned and remembered up until that day. At that point, students should have finished modules 1 through 4 and at least looked at modules 5 and 6. These tests could be a great indicator on how well students can perform during the real exam. These $\mathbf{Y}$ variables were tested using a Leave-one-out cross validation on an Extra Tree Regressor with the generated data set. As a reminder, scores are calculated with the Mean Absolute Error technique. Also, the results are also compared with a model that learns by simply using the median. This is to determine if the model has much merit. (Figure 4.7)




4.7.1: Boxplots of the error values

| | Score | Median Score | Score Difference | Real Median Grade | Median Est. Grade |
|---|---|---|---|---|---|
| QCL incidence white test | 3.44 | 3.37 | -0.08 | 13.33 | 12.321765 |
| QCL identification white test | 2.99 | 3.10 | 0.11 | 8.0 | 8.36 |
| Practical QCM white test | 2.58 | 2.57 | -0.004 | 10.36 | 10.29 |

4.7.2: Discrete Results




4.7.3: Actual grades compared to Predicated grades

Figure 5.7: Results from cross-validation of the White Tests

Unfortunately, the results are not great. The boxplots show very high median and quartiles. There are also a high number of extreme errors with some above 10. The scores represent the average error of the cross validation. With scores nearing nearing 3, in some cases, it's better to try and guess using the median value. For the incidence test, it seems that the scores are underestimated, this seems to be the case because there is a significantly high grades compared to low grades. There are many reasons why results are the way they are. This includes the fact that these grades do not rely on activity that occur after the exam date. When comparing grade the learned grades against the original grades, it's noticeable that the estimated grades tend to predict values close to the average of the actual grades. This makes it so that high and low actual grades when predicted tend to be more erroneous. The sample also lacks a good amount of extreme grades which gives the algorithm less to work with for these cases. Furthermore, the fact that the exams are ungraded puts less pressure on the students to succeed. Therefore, they might not take the tests as seriously as they could by not reviewing and studying the days before. (see Figure 3.8)

Since the models are made using regression trees, it it possible to study the features that had an impact in determining the predicted grades. Even though the results were not ideal, it's interesting to observe what images and variables had a big impact on determining the grades. These images can very well be the subject of a question on the exam. (Figures 4.8,

4.9, 4.10)

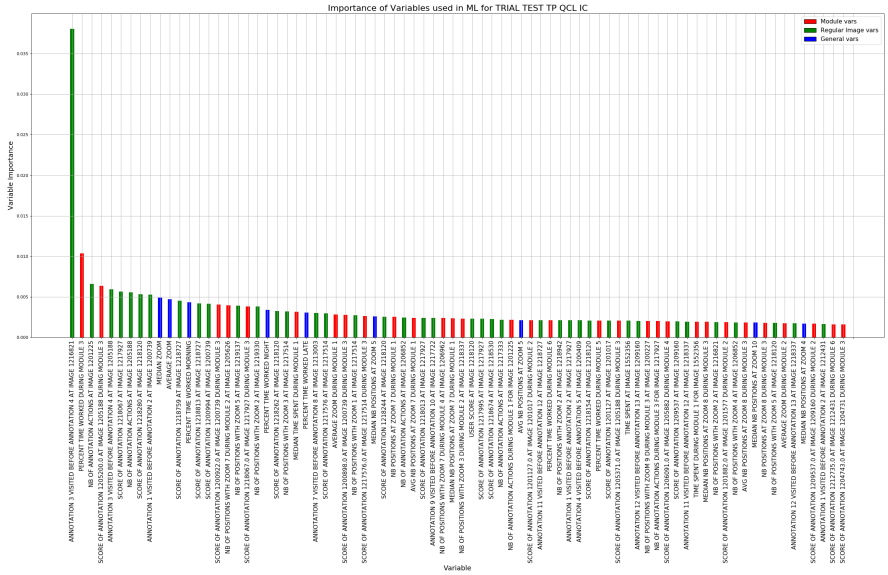



Figure 5.8: Feature Importance for QCL the Incidence

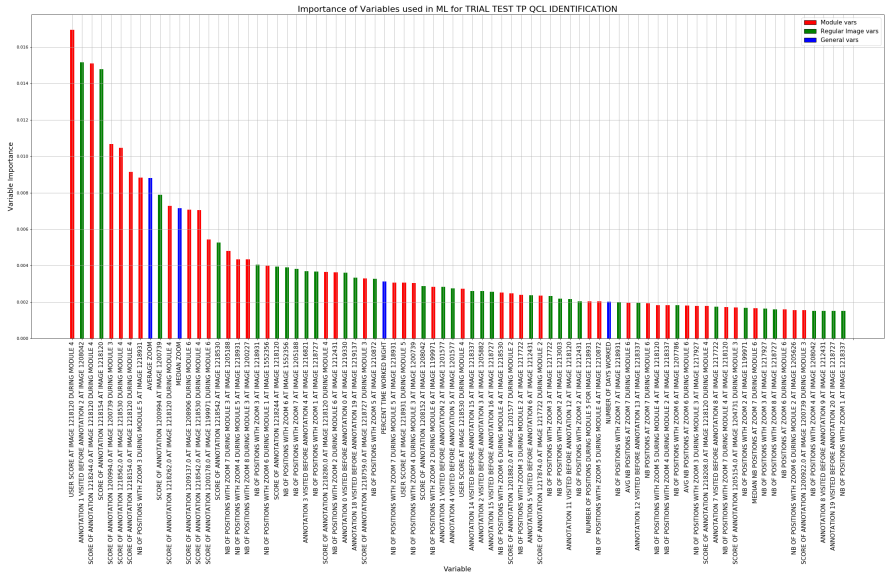



Figure 5.9: Feature Importance for QCL the Identification

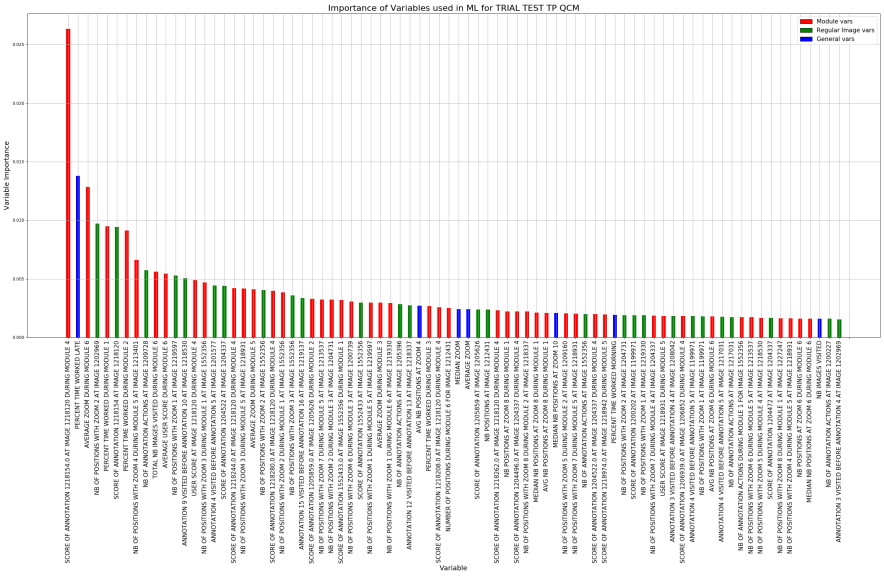



Figure 5.10: Feature Importance for the QCM

These figures show the top 80 features (or **X** variable) for their respective model. Even though each model shares the same features, their importance varies model to model. But it's probable that some features from a specific image can be at the top for multiple model. Anyways, features are split into 3 categories:

- **General Features :** Features that describes the data on the set of images as a whole. (average, median, etc..)
- **Regular Image Features :** Features that describe the data on a specific image. (number of positions at image XXXX)
- **Module Image Features :** Features similar to the two previous but associated to a specific module period. (number of positions at image XXXX during module Y)

Note that there are fewer general features, but many of those tend to have a high importance. Apart from the incidence test, it's appropriate to say that the module variables are the most impactful. This is because the white exam is taken in April instead of June. Therefore, user performance during specific modules are better indications on the performance during the white test. Also, for the identification test, the most impactful variable has a much higher importance than the rest. Even though this may look random, this variable may well be a good splitting factor. Unfortunately, this does not show the overall impact of certain images. (Figures 4.11, 4.12, 4.13)

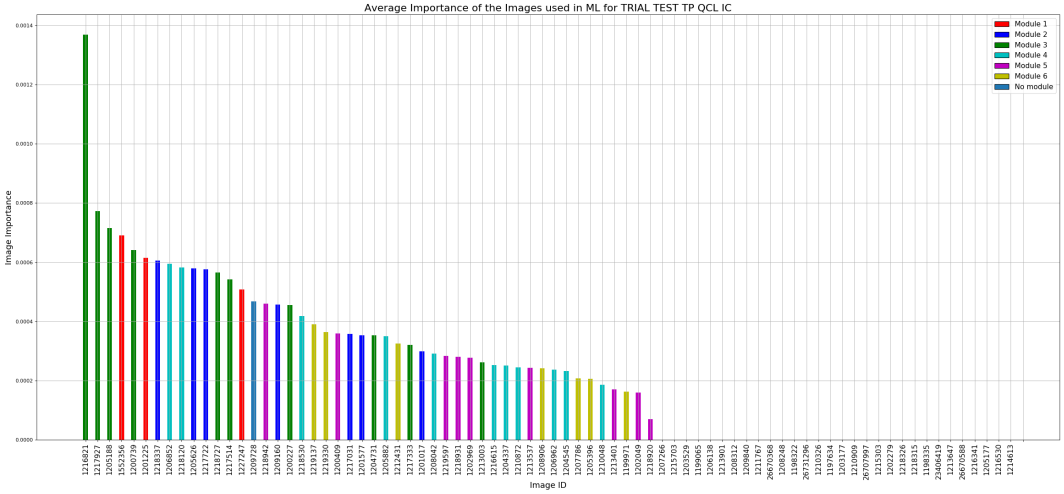



Figure 5.11: Image Importance for QCL the Incidence

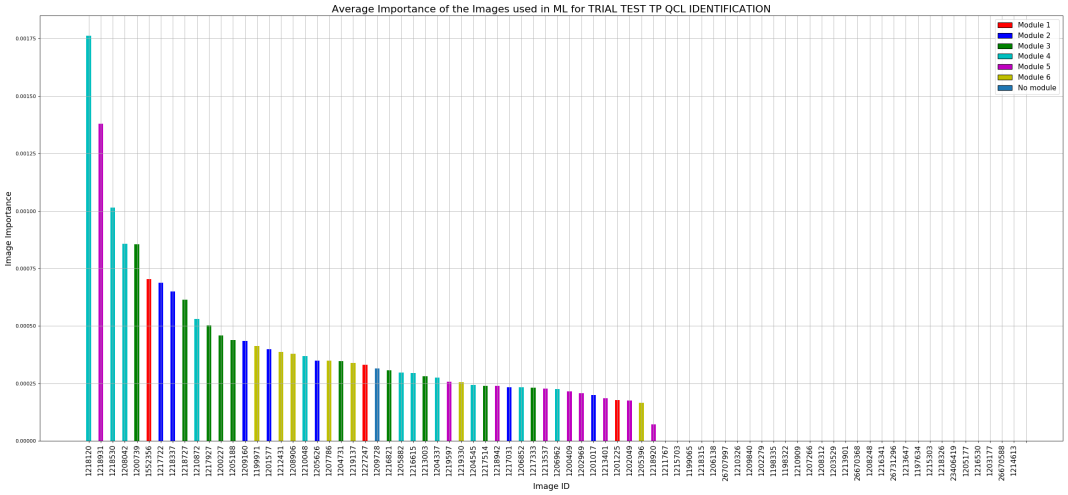



Figure 5.12: Image Importance for the QCL Identification

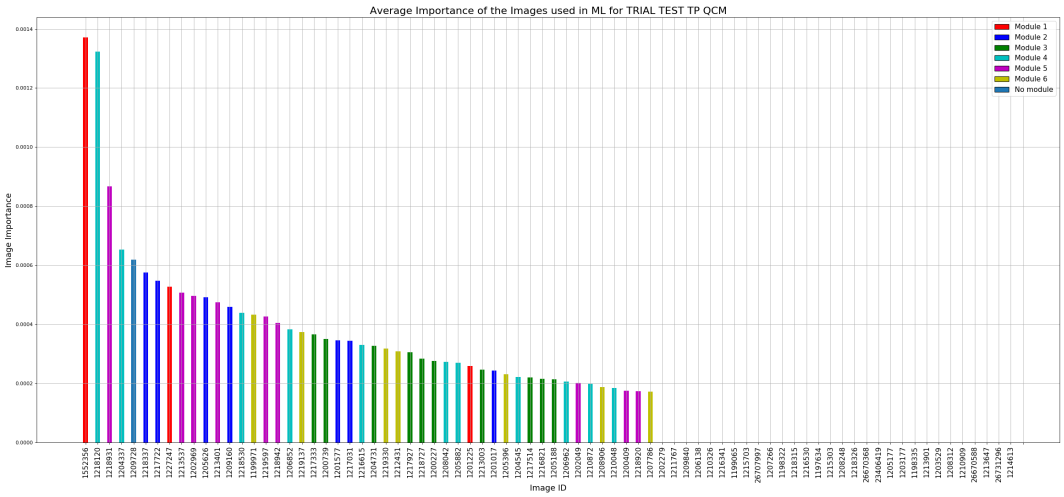
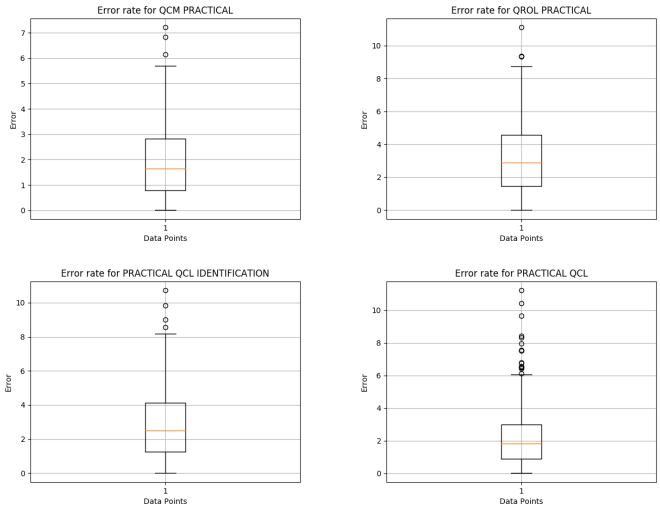



Figure 5.13: Image Importance for the QCM

As a reminder, out of all the images in the GOLD project, about a third of the images were not even visited. Also, some of the images don't belong to any module. It seems that as a general consensus that the most impactful images do not belong to the last module. This is due to the fact that the module 6 is not included in the trial exam. It's also worth noting that the module 1 is not included in the exam either. But since the module 1 is considered a tutorial, it can be useful. An image that appears often at the top is the image 1218120. This image was used for a preparation identification exercise. It seems that those who perform better in regards to that image on Cytomine do better in the white tests. For the QCL identification, the exercise images are often the most important over all the images in the same module. This seems to prove that doing the exercises seriously and correctly can lead to better grades. For the QCM, the image 1552356 which is a module 1 tutorial image is the most impactful. Those who follow the tutorial may have an easier time understanding the course from the start and may perform better for the QCM. As for the incidence test, the module 3 variables seem to be the most impactful. This module may contain examples that allow the student to better differentiate different incidence angles for specific objects in the images. OF course the cross validation results are not the best, so these assumption are not necessary correct. This leads to the actual practical exam.

### 5.3.3 Practical Exam Grades

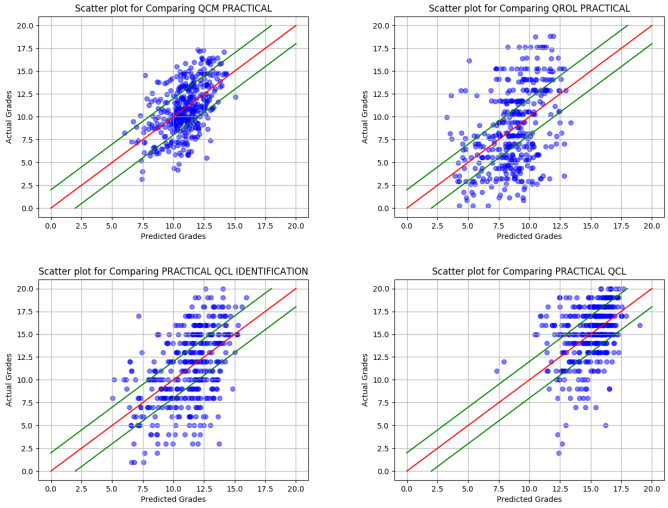
There are a total of 5 practical exam results including the total. These **Y** variables tested similarly to the white tests. The exam took place on the 12th of June 2017. Unlike the white exam, the students should have finished working on all the modules. These exams also depend on the entirety of the program unlike the white tests. Since these exams are graded, students should have the motivation to study and apply themselves. (Figure 4.14)




4.14.1: Boxplots of the error values

| | Score | Median Score | Score Difference | Real Median Grade | Median Est. Grade |
|---|---|---|---|---|---|
| QCM practical | 1.94 | 2.29 | 0.36 | 10.80 | 11.01 |
| QROL practical | 3.23 | 3.64 | 0.40 | 7.93 | 8.80 |
| QCL identification Practical | 2.86 | 3.36 | 0.5 | 11.0 | 11.61 |
| QCL incidence Practical | 2.23 | 2.83 | 0.59 | 15.0 | 15.32 |

4.14.2: Discrete Results




4.14.3: Actual grades compared to Predicated grades

Figure 5.14: Results from cross-validation of the White Tests

the results obtained are more interesting. The results vary heavily depending on the exams that were taken. All the scores are significantly better than their respective median compared to the trial tests.

- **QCM practical :** The best score, the boxplots show the most well rounded results. The 75 percentile of grades estimated have an error of less than 3 and a 25 percentile of grades with an error of less than 1. Unfortunately, there are still some cases with a significantly high error. As for relation between the expected grade and the estimated grade, there is a clear pattern following the red line. The exception being that low actual grades are overestimated by the model. QCM exams are usually are straight to the point. Students need to know the answer to question but they don't have to explain it. All students by definition are graded equally because the exam is straightforward and easy to correct.

- **QROL practical :** As opposed to the QCM, the score is significantly worse. This is also shown by the boxplot where the 75 percentile is at less than 4.5. By comparing the grades, it's apparent that there are more extreme grades and their estimations are off for the most part. This is normal because students write long and detailed responses to the questions. The teachers are more critical of the students. They look to see if students understand the contents of the course as opposed to learning by heart.

- **QCL identification Practical :** With a score of 2.86, this QCL does not offer the best results. It follows a somewhat lessened pattern of the QROL. As this portion of the exam gives the student an exhaustive list of possible answers, it's hard to a student to guess. It's critical that the students knows what they observe. Similarly to the QROL, lack of certainty lowers the grade. As opposed to the QCM where uncertainty is not much of an issue due to the lack of options.

- **QCL incidence Practical :** the incidence score is somewhat better. Apart from a couple exceptions most of the real grades have a small variance. This helps explain the very low 75 percentile of about 3 and the big number of erroneous cases above 6. Since this part of the exam is similar to a QCM but with only 3 options, it's relatively easy to answer the questions correctly. This explains the higher grades for most students.

When taking practical exams the students have images to look at. Certain aspects of the images need to be identified before answering the questions. Unfortunately with the data fetched from Cytomine, it's hard to determine whether or not the students were able to identify certain concepts when using Cytomine. But attempting to identify the features that are close to determining whether or not a student understood the course is a start. (Figures 4.15, 4.16, 4.17,4.18)

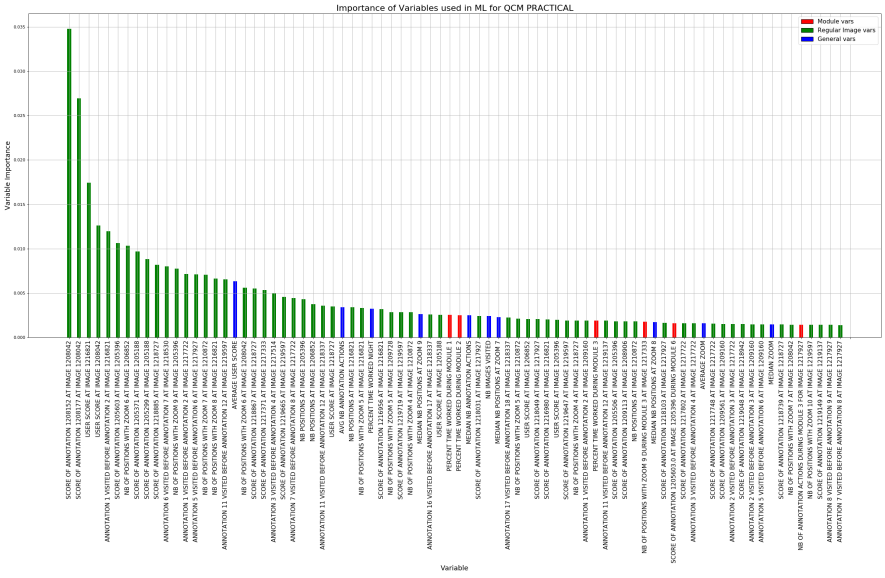



Figure 5.15: Feature Importance for the QCM Practical

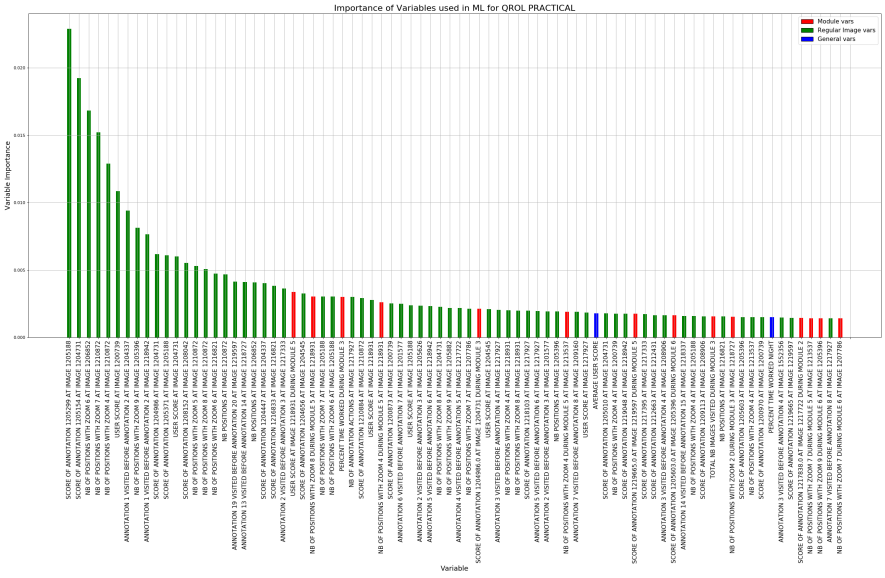



Figure 5.16: Feature Importance for the QROL Practical

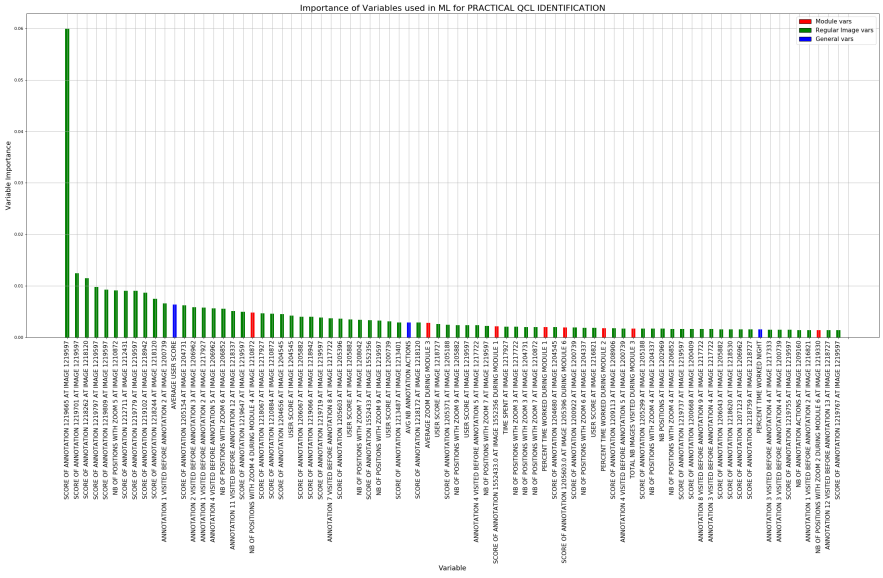



Figure 5.17: Feature Importance for the QCL Identification

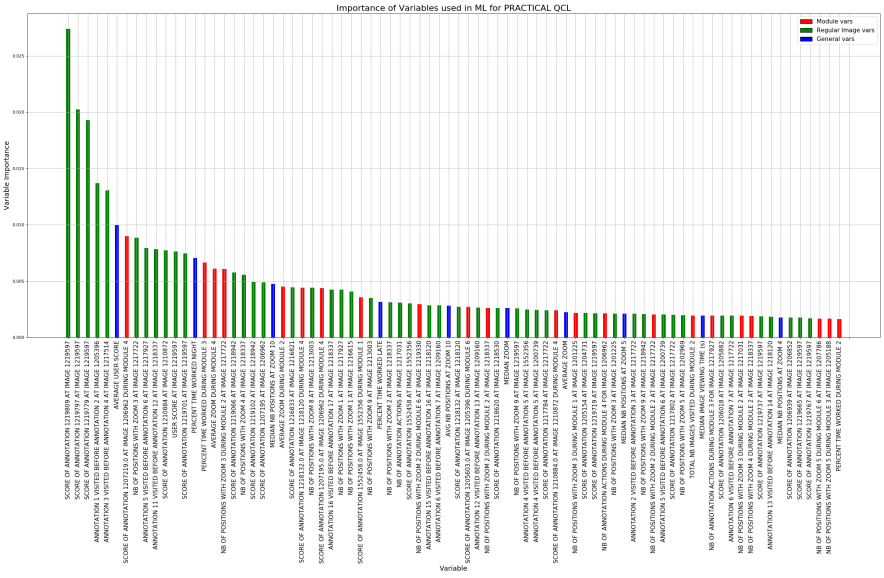



Figure 5.18: Feature Importance for the QCL Incidence

The results are different from the trial tests. For one, the Module features are relatively less impactful than before. This shows that the total work done throughout the semester is more important than working more during specific modules. It also is shown that the students spent the most time on Cytomine the day before the exam (see Figure 3.8). Similar to the white tests models, the Average Score feature always shows up as an important feature. This variable seems to be the closest variable into determining whether or not a

student understands the course but it's still not perfect. In the end, some images and some modules have more impact than others. (Figures 4.19, 4.20, 4.21,4.22)

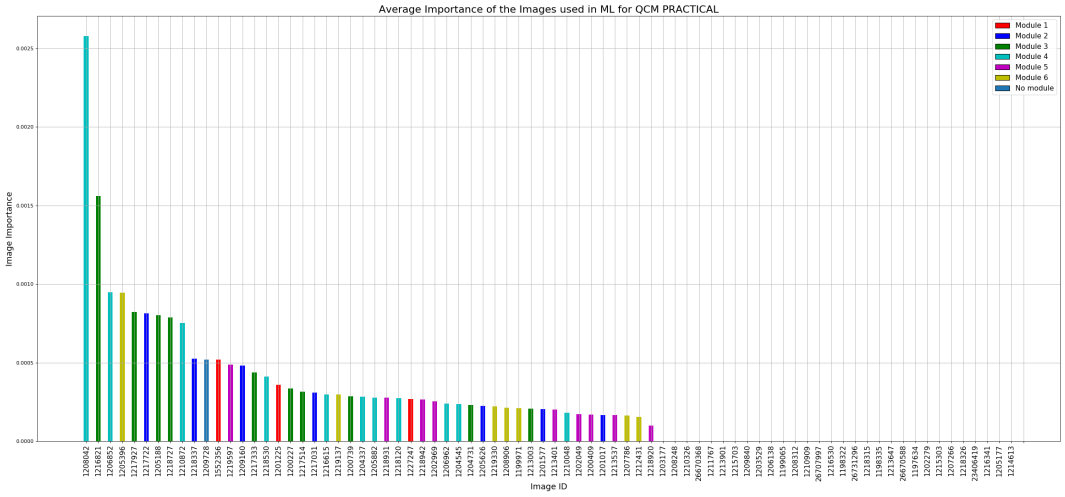



Figure 5.19: Image Importance for the QCM Practical

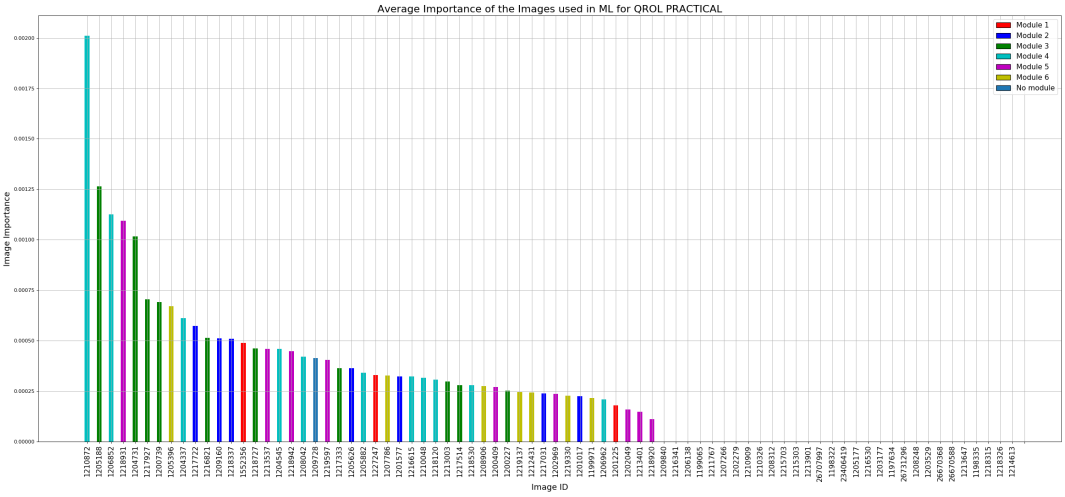



Figure 5.20: Image Importance for the QROL Practical

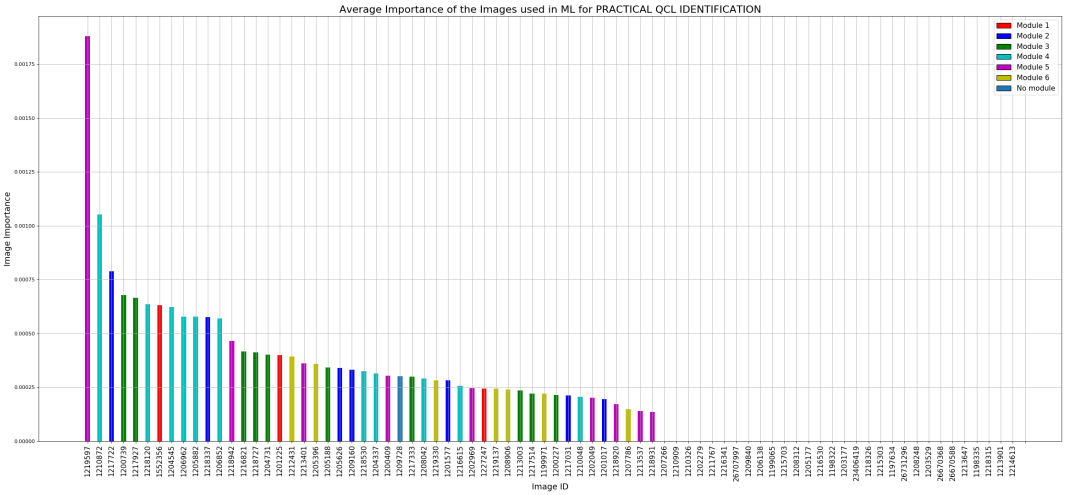



Figure 5.21: Image Importance for the QCL Identification

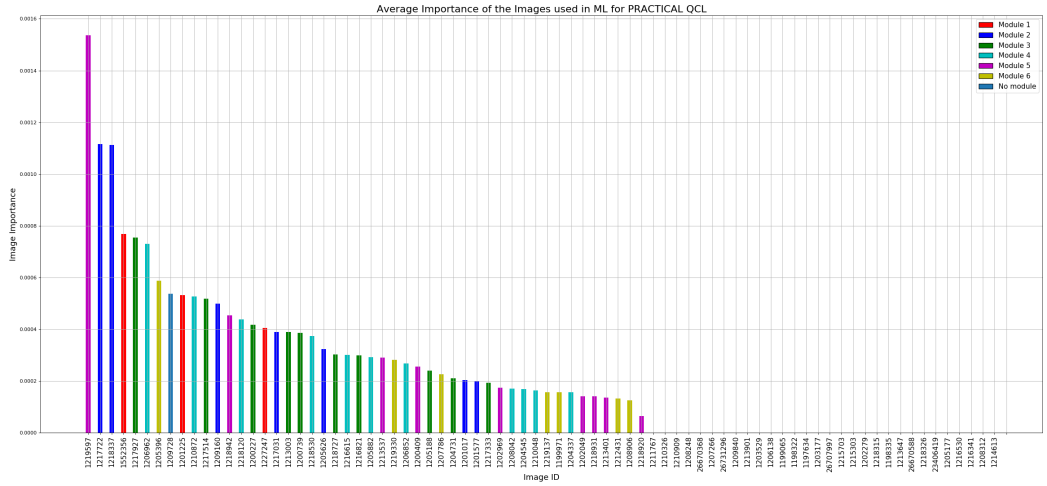



Figure 5.22: Image Importance for the QCL Incidence

It's hard to determine which modules were are more impactful as a whole. It seems that each module has at least one image that has a big impact on the result for each test. Like before, most these images are the ones dedicated to exercises. This supports the hypothesis that doing the exercises has a positive impact on the grades. But for example the image 1218120 does not appear at the top anymore. This shows that even though the models follow the same patterns, it's still varies based on the contents of the exam.

Sometimes it's interesting to look for correlations between the expected result and some features. This helps to see its direct impact in estimating a grade. In fact, with most of the the features, a higher value should mean a better score. To test this, the Pearson Correlation was calculated for the top 6 features of the Practical QCM. (Figure 4.23)

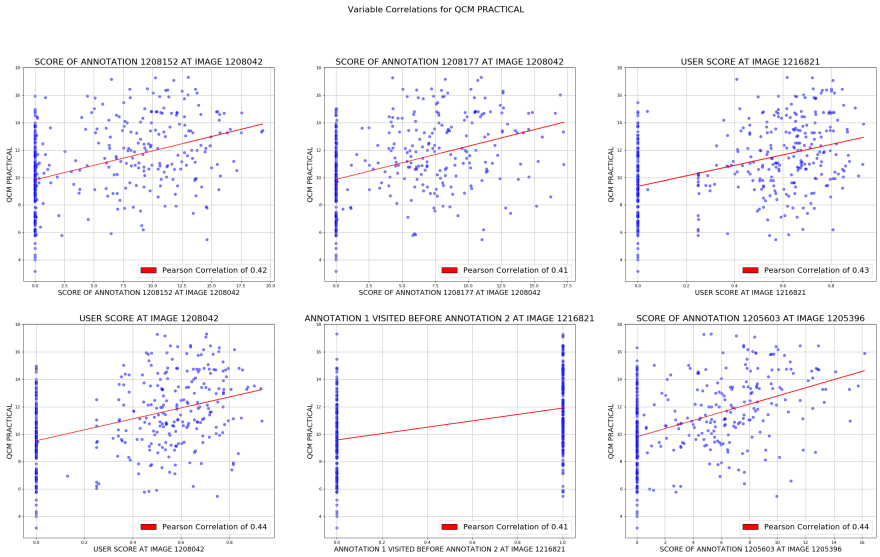


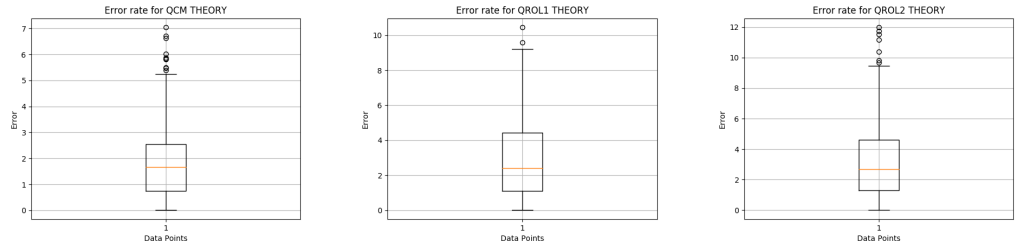

Figure 5.23: Pearson Correlation

Most of these features are score features which in short tells the model how well the user viewed the image in regards to the annotations. Therefore a higher value means a better observation of all the annotations. This correlation means that the more time the student is focused on annotations, the more likely that student is to get a better grade. This could explain the somewhat good Pearson Correlation for these variables.

After studying the results for the practical exams, a question that can be asked is whether or not this analysis can work on Theoretical Exam grades.

### 5.3.4 Theoretical Exam Grades

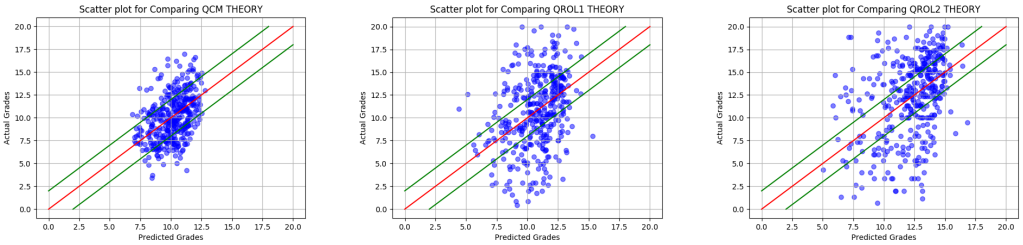
The same tests were launched for the theoretical grade results. The main difference for the theory portion is that there are no QCL tests and two QROL tests. The learning done on the QROL test results of the practical showed the worst results. Meanwhile the QCM provided the best results. This trend is likely to continue. (Figure 4.24)




4.24.1: Boxplots of the error values

| | Score | Median Score | Score Difference | Real Median Grade | Median Est. Grade |
|---|---|---|---|---|---|
| QCM Theory | 1.86 | 2.05 | 0.19 | 10.03 | 10.24 |
| QROL1 Theory | 3.00 | 3.30 | 0.30 | 10.87 | 11.05 |
| QROL2 Theory | 3.25 | 3.60 | 0.36 | 13.0 | 12.67 |

4.24.2: Discrete Results




4.24.3: Actual grades compared to Predicated grades

Figure 5.24: Results from cross-validation of the practical tests

As predicted, the QCM yielded the best scores while the QROL provided sub par scores. The assumptions made based off of the practical exams were correct. The Extra Trees has an easier time predicting QCM exams. But in this case, the QCM grades has a lower variance then the QROL.

The following figures show the feature importance for their respective model (Figures 4.25, 4.26, and 4.27).

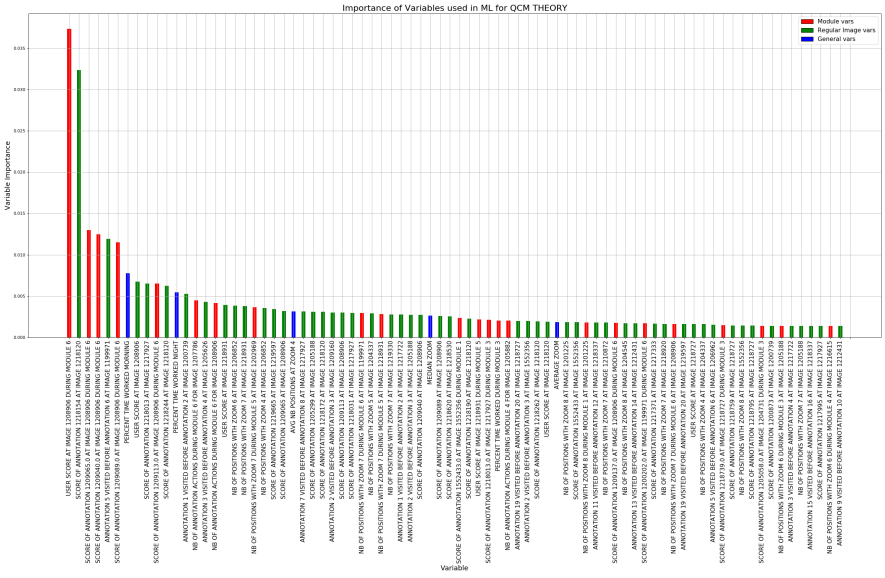



Figure 5.25: Feature Importance for the QCM Theory

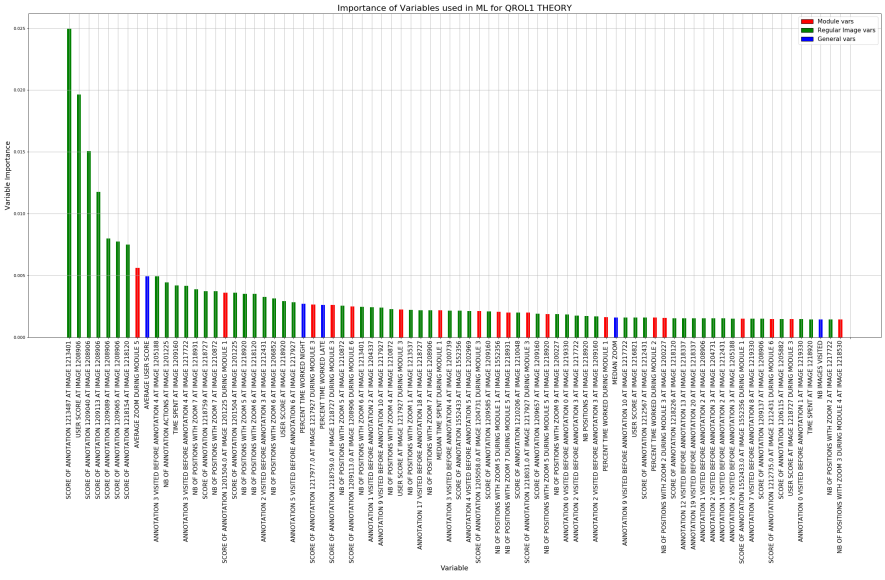



Figure 5.26: Feature Importance for the QROL1 Theory

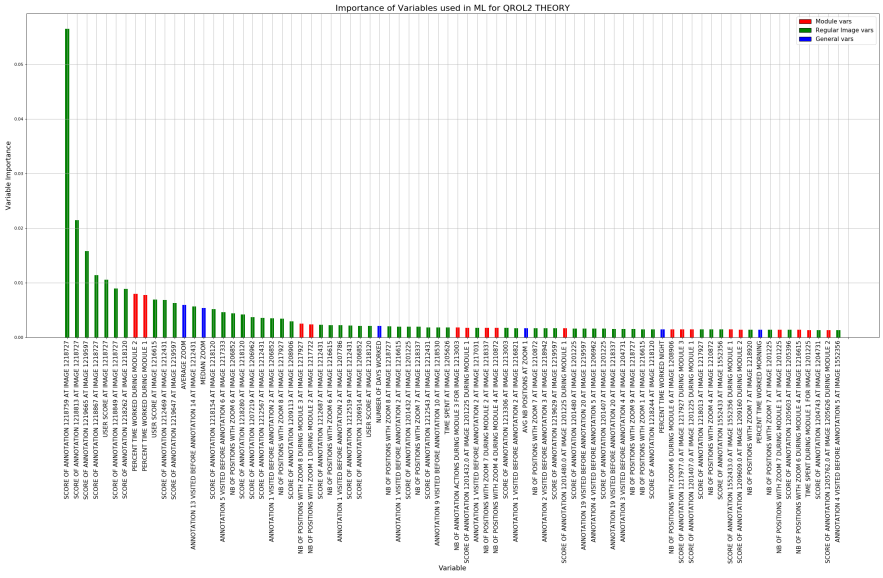



Figure 5.27: Feature Importance for the QROL2 Theory

As usual, the scores are the most prominent features for all the theoretical exercises. It's interesting to note some differences with the practical exercices. This includes features that are associated to the time worked. It would not be too much of a stretch to theorize that studying with Cytomine during certain time periods, the student would assimilate more information. For example, when working after 2AM, the students may be more tired and they might retain less information.

As for images themselves, the QROL questions are more specific so certain images may be the centerpiece. (Figures 4.28, 4.29, and 4.30)

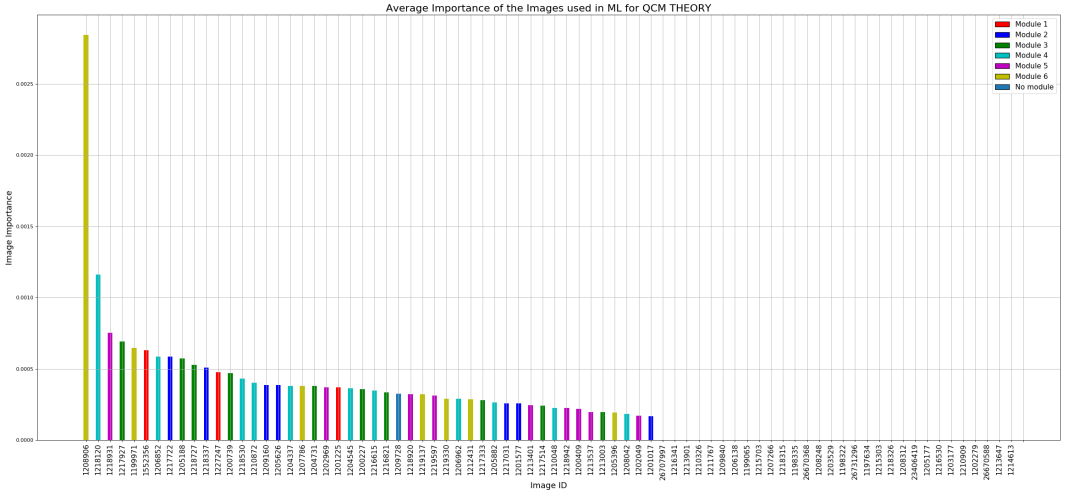



Figure 5.28: Image Importance for the QCM Theory

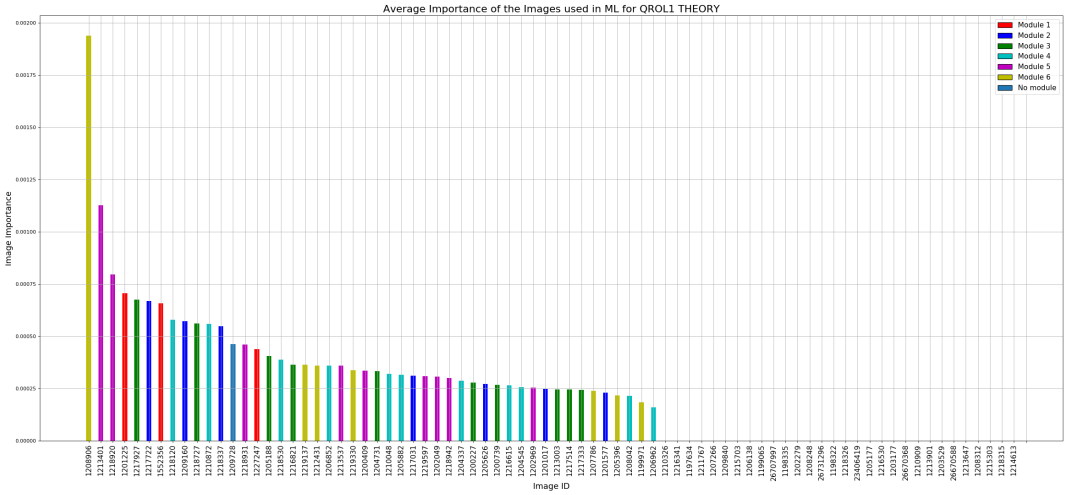



Figure 5.29: Image Importance for the QROL1 Theory

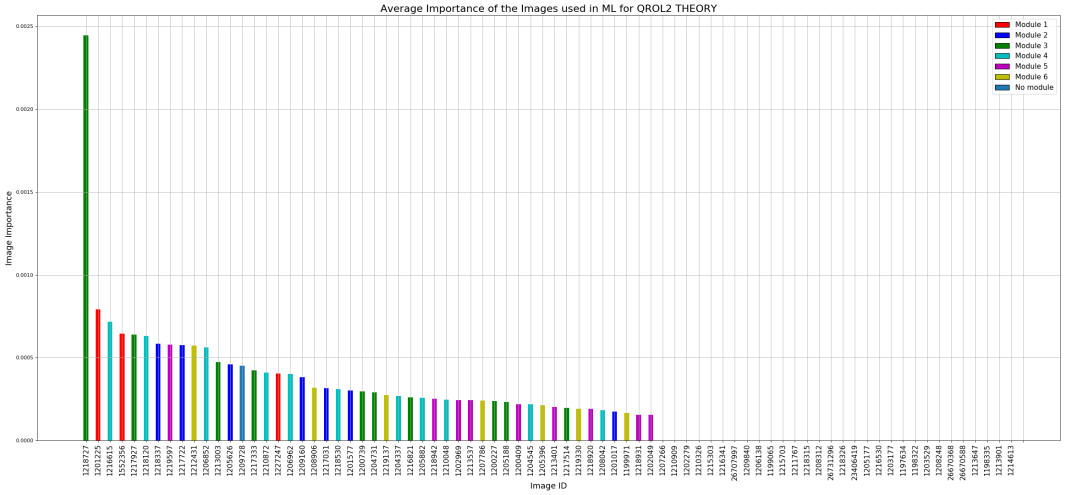
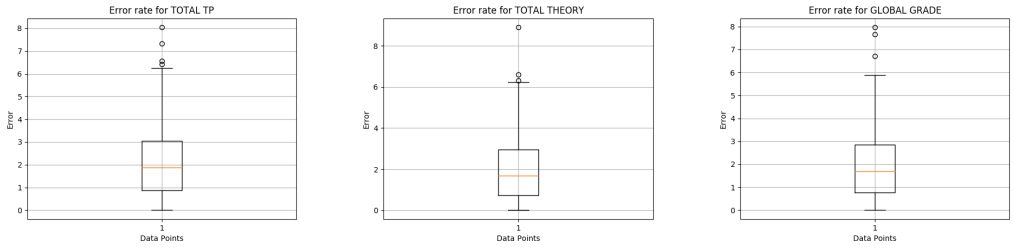



Figure 5.30: Image Importance for the QROL2 Theory

As predicted, some images and modules are more relevant then others. For the QROL1, it seems that the question was associated information that the students learned from the modules 5 and 6. The image 1208906 which is an image used for learning is the most important. This is contrasted by the QROL1 question where the more important images are associated to the earlier modules (1 and 3). In this case, the image 1218727 was the most impactful. This image was given more as an exercise image as opposed to the previous one. As for the QCM, the image 1208906 also stood out. It's possible that, there's more educational interest behind this image.

### 5.3.5 Global Grades

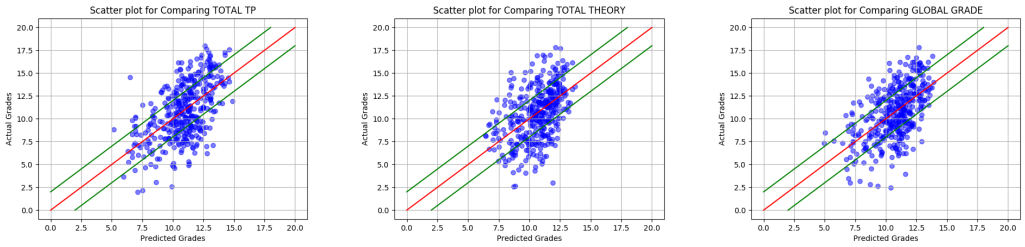
The next objective is to study these grades as a whole. This includes the total theoretical grades, the total practical grade, and the exam as a whole. (Figure 4.31)




4.31.1: Boxplots of the error values

| | Score | Median Score | Score Difference | Real Median Grade | Median Est. Grade |
|---|---|---|---|---|---|
| Total Practical | 2.12 | 2.59 | 0.47 | 10.80 | 11.06 |
| Total Theory | 2.03 | 2.34 | 0.31 | 10.79 | 10.86 |
| Global Grade | 1.95 | 2.37 | 0.42 | 10.72 | 10.95 |

4.31.2: Discrete Results




4.31.3: Actual grades compared to Predicated grades

Figure 5.31: Results from cross-validation of the exam results

As a whole, there's a mean absolute error of about 2. Overall, still not the best results. An interesting observation is that the learning algorithm has a easier time predicting theoretical grades over practical grades. As using Cytomine is overall a practical activity, the fact predicting theoretical grades works just as well can possibly imply that students can retain theoretical concepts with Cytomine.

Observing what behaviors lead to this grade can prove useful : (Figures 4.32, 4.33, and 4.34)

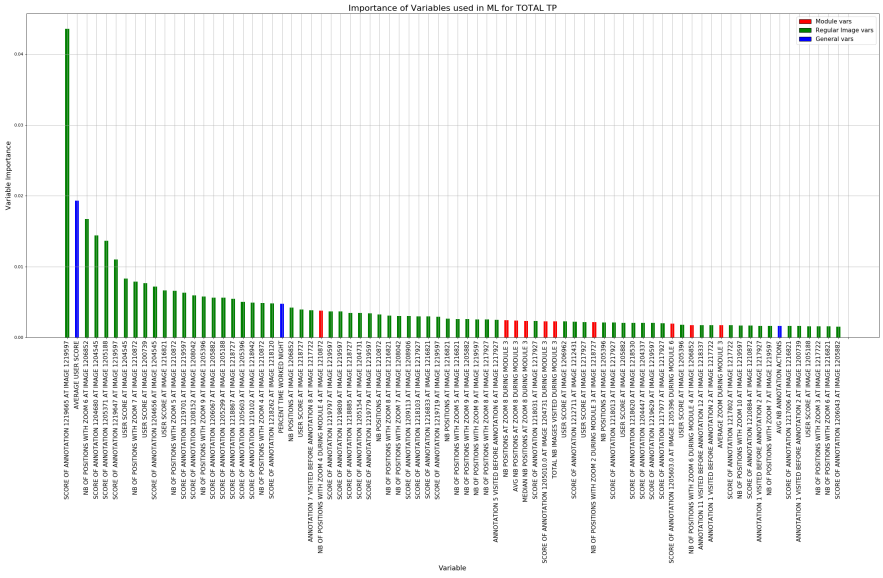



Figure 5.32: Feature Importance for the total Practical

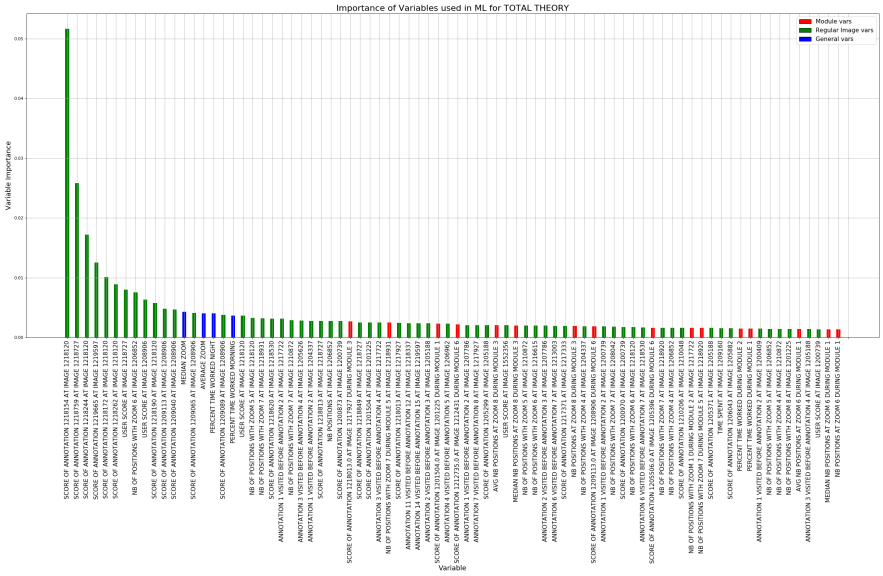



Figure 5.33: Feature Importance for the total Theory

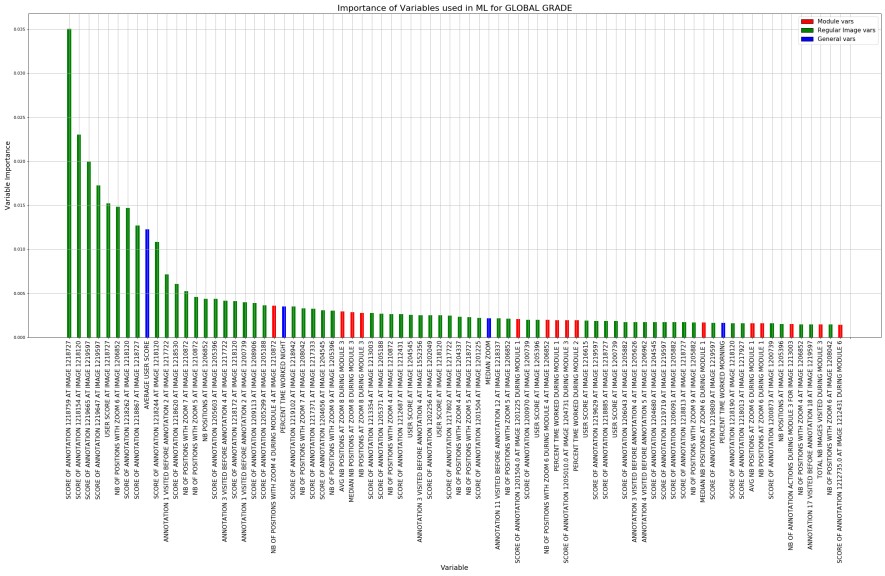



Figure 5.34: Feature Importance for the global Grade

Overall, the most impactful features are often associated to the students' performance scores for all the models. This indicator that the student followed the instructions has a notable impact on the results. Also, module features are also much less impactful than their regular counterparts.

As for the images themselves the results are : (Figures 4.35, 4.36, and 4.37)

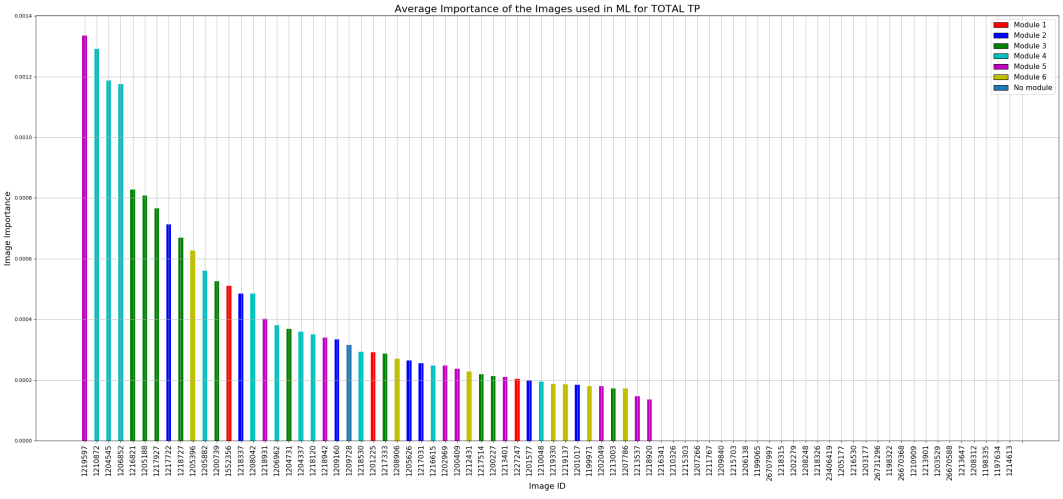



Figure 5.35: Image Importance for the total Practical

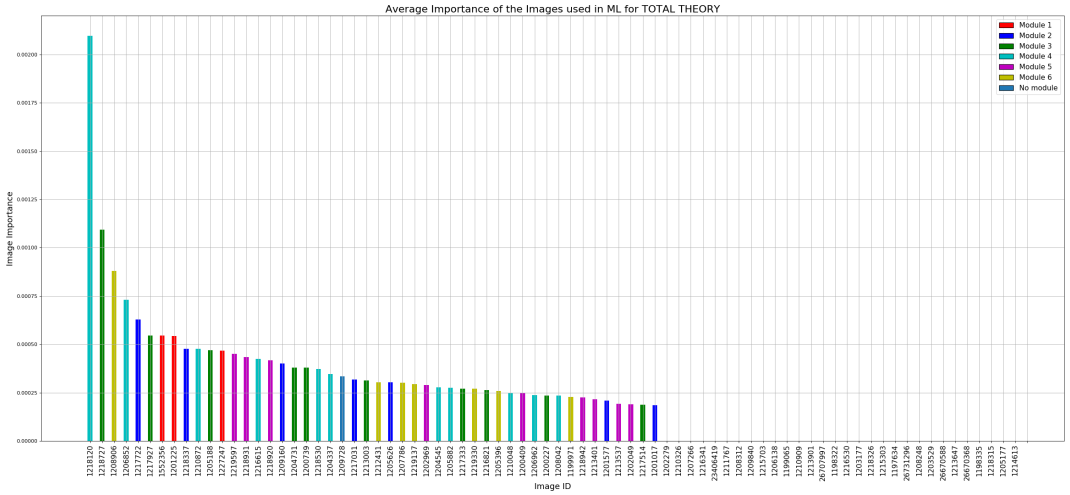



Figure 5.36: Image Importance for the total Theory

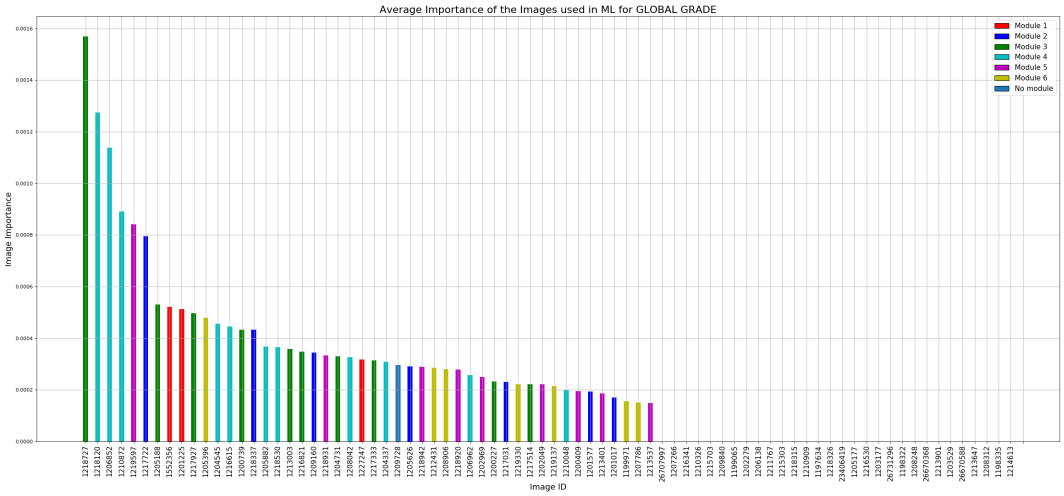



Figure 5.37: Image Importance for the global Grade

For the most part, the important images are similar to those of all the individual tests put together. So as a whole, there's a sizable set of images that have a high impact on the grades predicted. These results give an indication on what patterns can lead a student to pass.

### 5.3.6 Learning with additional Information

A small experiment is to set the White Test grades as features and learn the Global grades. This extra information can prove usefull. The experiment is to see if those who did well on the white tests do well on the main exam. For those who did not participate in the trial tests, they receive a value of -1 for those features. The results are : (Figure 4.38 & 4.39)

- Score : 1.79
- Score from Median Model : 2.37
- Score Difference : 0.58
- P-value : 0.92
- P-Value of Median Model : 0.67

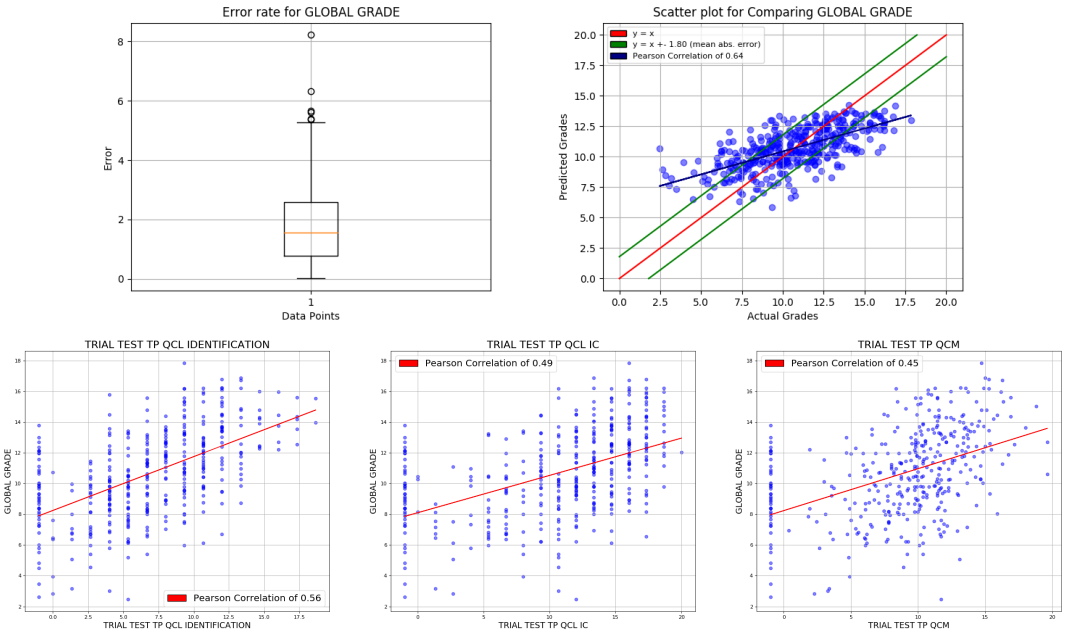



Figure 5.38: White Test as features results - Statistics


Figure 5.39: White Test as features results - Feature Importance

As is clearly visible, The three white tests had an impact on the results. They became by far the most important features according to Figure 4.39. There's also a strong correlation between those features and the exam results. It's hypothetically possible that in general, students who do better for the white tests do better on the final exam. This extra information also helped predict better results as a whole. This explains the reason why Teachers heavily encourage students to take the white tests. It gives student an idea on how the exam will be like. They can also get feedback if they had any questions. Of course, the cross validation results are still far from ideal so it's still dangerous to make too many assumptions.

The next experiment tries to learn practical results all while knowing the theoretical results.

TODO

# 6 Discussion

## 6.1 Negatives

First of all, the results obtained from the Machine Learning were subpar. A mean Absolute error of around 2.0 out of 20 is 10% of the range which is not great. The main cause of which is the small data set of 395 students. If it was possible to group people up based on habits and patterns, 395 students would not be enough for such study. It's still important to stay skeptical because every person is unique. Some people have a much easier time learning than others. In such cases, the student may not need much time to understand the concepts provided by Cytomine. Therefore, there there are less recorded positions and actions. In other case, the exact opposite. When on Cytomine, the analyst can't be in the person's head when using the application. The student can go on the website and retain nothing.

Also, the Gaze data may not be perfectly accurate. Without sensors that observe the user's eye movements, it's very impractical to track Gaze data. In most cases, a person is usually focused on or near the center of the screen. It's possible that they can focus more towards a corner. In it's current state, corners have a lower value from the calculated Gaussian distribution. It's also worth to mention again that positions are still recorded even when the window is not active. Since this issue is not easy to resolve without losing information, it also may lead to some inaccuracies. It's also important to note that much information is contained in the course itself. The course contains numerous examples of cells and tissues that is also found on Cytomine. It's entirely possible that students can ignore Cytomine for the most part and just learn from the course's Syllabus.

But overall it may be a good thing that the results were not so great. It is important to keep in mind that this study is not used to encourage students to follow the directions to the letter. The students should be free to work on their course the way they see fit. If they were to fail, it is their own problem. It's also important to encourage different styles and methods, and not force a particular way of thinking. This is why the Annotation visit order features were not very impactful. If those variables were more impactful and the results were better, it could have implied that visiting annotations in a specific order would be important to passing the exam. Overall the goal is not to try and dictate how the students work.

Secondly, even with sizable number of features it still feels like it's still possible to infer new features. The problem is to find features that can have a real impact on the results. For example, it's possible to generate the same features but for each day of the semester. This would end up adding tens of thousands of features. Due to the small number of individuals, it would be hard to draw conclusion for many of these features. It would also be just as hard to determine whether or not students understand the course.

Thirdly, Cytomine comes with tools to create and edit annotations. Unfortunately, in its current state annotation operations are only accessible to the project manager. It would be interesting for students to be able to create personal annotations. There could be multiple applications, including :

- For exercises, students can be asked to look for certain objects in images. These objects can be annotated. These annotations can be seen by the teachers. After a certain deadline, teachers can correct and give advice to students.
- For the images used for teaching, students can annotate and describe certain objects similarly to before. But this time, it's used to ask questions to the teachers. For example on Cytomine, teachers can be notified. Afterwards, they can reply to these questions. This can be a interactive way fo students to communicate with teachers.
- Private annotations for the students, Simply to keep information on certain objects saved and easily accessible.

This option for students can help improve the course in the long run.

### 6.2 Positives

This tool was able to generate many ways to visually observe user behavior on Cytomine. This includes :

- Gazemaps (Gaze Heatmaps)
- Scan Paths
- Timelines
- Feature importance graphs
- Image Importance graphs
- Feature correlation graphs
- Cross Validation error Boxplots
- Cross Validation error Scatter Plots

These tools can give indication on what the students are doing and when they are doing it. It's unfortunate that generating some of the figures may take some time. But with this tool, it's possible to generate Scan Paths or Gazemaps for specific image and user pairs on demand. This can be directly implemented on a Cytomine website and not necessarily only for the MOOC.

Even though the Machine Learning results were not the best, the variable importance can still give small indications on what can lead students to pass or fail exams. For example, the features that are student's performance score in regards to annotations quantifies how well a student follows instructions. These features often are the top in regards to importance. The main thing that can be determined is which annotations had the most impact for the exam. This leads to image importance, and which images were the most impactful. A good example is for the QROL1 and the QROL2, the differences are clear. One question was most likely more associated to the earlier portion of the course while the other was more associated to the end of the course. This is theorized without actually knowing the contents of the two questions.

## 7 Conclusion

Overall, there were some interesting results. This tool provided the means to analyze and visualize user behavior on Cytomine. Gazemaps and Scan paths provide some key information on the viewing pattern for a single image while timelines gave an overview on the project as a whole. Machine Learning methods have been attempted and the results give much insight on groups of people but also individuals. The results obtained were not the best due to the lack of individuals and the fact that this study has been carried out on people who each has there own way of thinking and doing things. The goal was not to turn people into robots where there is only one good way of doing an activity, it's more to understand the reasoning behind decisions and why it was done. Of course there still is much more

information that can still be extracted and there may be even more complexity associated to it. Without a background in Psychology, it's a complicated task to understand what's going on in a person's head when they use Cytomine. Anyways this study proposes new tools that be used by Cytomine to draw statistics and even some minor adjustments to the MOOC.

# 8 References


- [ Maree R, Rollus L, Stevens B, Hoyoux R, Louppe G, Vandaele R, Begon JM, Kainz P, Geurts P, Wehenkel L.] Collaborative analysis of multi-gigapixel imaging data using cytomine. Bioinformatics (2016) 32 (9): 1395-1401.
- [ Dalmaijer, E.S., Mathôt, S., & Van der Stigchel, S. ] PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eye tracking experiments. Behaviour Research Methods, http://www.pygaze.org/2015/06/pygaze-analyser/, https://github.com/esdalmaijer/PyGazeAnalyser, doi:10.3758/s13428-013-0422-2.
- [ http://scikit-learn.org/, https://arxiv.org/abs/1201.0490 ]
- [ Geurts et al. 2006 DOI: https://doi.org/10.1007/s10994-006-6226-1 ]
- [ Multon S, Weatherspoon A, Schaffer P, Quatresooz P, Defaweux V. ] Practical histology in tune with the times. Med Educ 2015, 49(11):1166-7.
- [ Dmitriy Shin, Mikhail Kovalenko, Ilker Ersoy, Yu Li, Donald Doll, Chi-Ren Shyu, Richard Hammer ] PathEdEx Uncovering HighâĂŚexplanatory Visual Diagnostics Heuristics Using Digital Pathology and Multiscale Gaze Data. J Pathol Inform 2017.
- [ lawomir Walkowski1, Mikael Lundin, Janusz Szymas, Johan Lundin ] Exploring viewing behavior data from whole slide images to predict correctness of studentsâĂŹ answers during practical exams in oral pathology. J Pathol Inform 2015.
- [ Daniel Santiago, Germán Corredor; Eduardo Romero] A sparse representation of the pathologist's interaction with whole slide images to improve the assigned relevance of regions of interest. 2017.