Towards image-based cancer signatures from histopathology data

First year report

Naylor Peter
Supervised by: T. Walter and F. Reyal
Units: Center for Computational Biology and UMR900

8th of June 2016

Outline

Outline

Introduction

Image data in biology or in medicine can be acquired in these settings:

- Fundamental research, through home brewed experiments:
  - Microscopy data of cell cultures.
  - In High content screening, many films (thousands) of experiments undergoing a different change.
- In clinical practice: through practice of general medicine:
  - Tissue scans, such as histopathology slide (PhD project)
  - Photographs
  - MRI scans
  - Ultrasound

  Etc

Outline

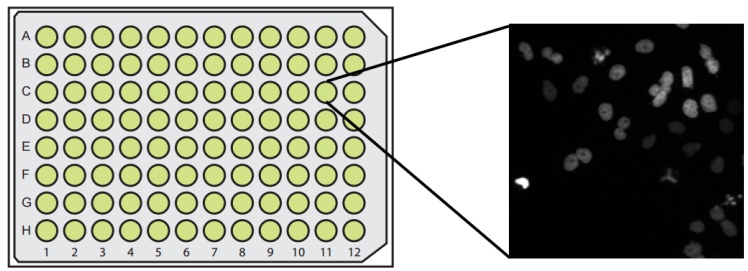
Time-lapse fluorescent microscopy data in the context of High Content Screening

Time-lapse microscopy: Microscope image sequences, gives an accelerated view of the microscopic process.

Fluorescent: In our case, after altering the experimental cells, certain proteins within the cell emit visible wave length. These proteins are used to highlight specific structures.

- H2B-eGFP is informative about the mitotic events.
- PCNA-mcherry is informative about non mitotic events.

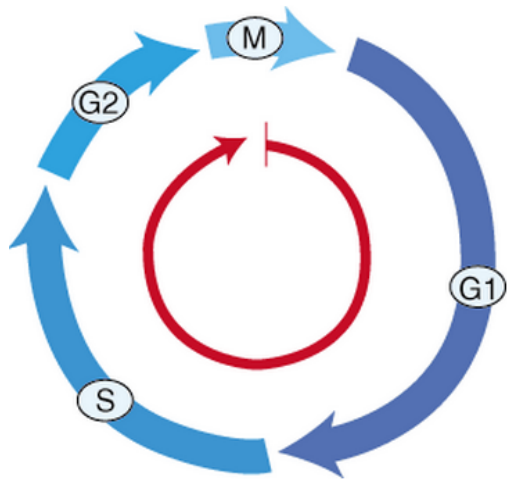
Incenp



High Content Screening

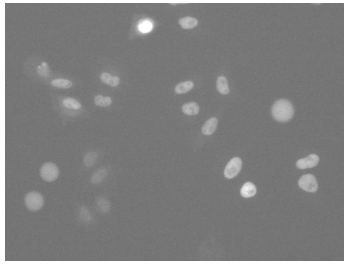
Recycling the MitoCheck project

The mitocheck dataset is a unique dataset with a chromosome marker (H2B-eGFP). 200 000 filmed loss of function experiments. Can we used this data set to study the non-mitotic cell cycle phases? Idealy we would use a replication marker (PCNA-mcherry) but it is absent here.



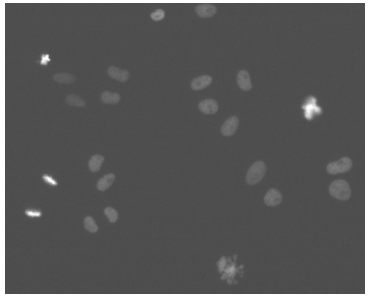
Cell cycle

The raw data



Data acquired by Michael Olma with the PCNA marker

- Data set that helps us label the data.
- HeLa cells stably expressing PCNA marker which is informative about the non-mitotic phases.

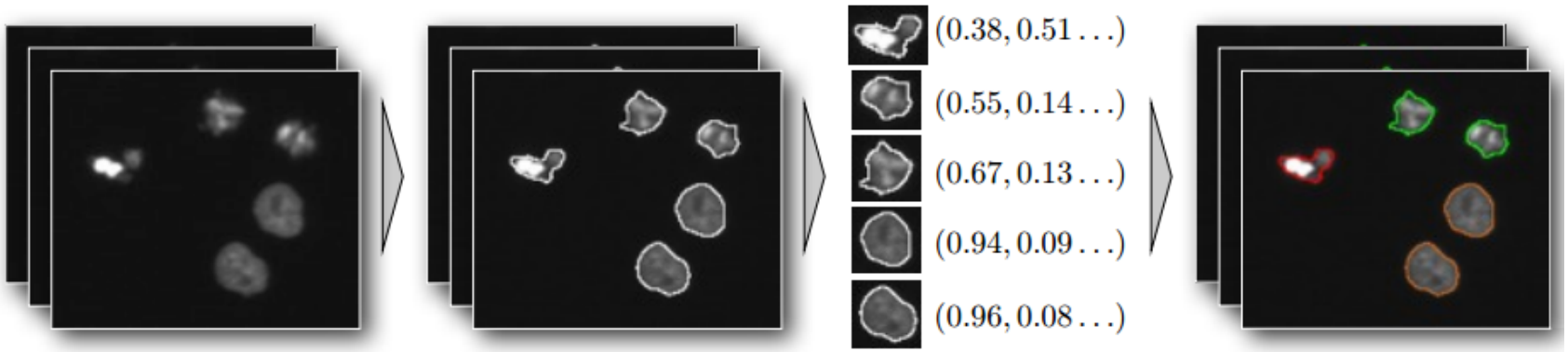


Data acquired by Michael Olma with the H2B marker

- Data set from which we extract features for training purposes.
- Looking for paterns that help differenciate non-mitotic phases.
- HeLa cells stably expressing H2B marker, informative about the mitotic phases.

Extracting the data

Using *CellCognition*:



CellCognition steps for feature extraction

Results on the cell cycle phases length

To assess our predicition, we looked at different cell cycle phases length out of 234 trajectories:

Length of:	Mean	Standard deviation	Number of trajectories
G1	7.13	3.8	159
S	6.28	2.9	124
G2	3.17	1.4	124
Cell Cycle	16.7	4.1	124

Table : On Michael Olma set with the PCNA channel, our "ground truth"

Length of:	Mean	Standard deviation	Number of trajectories
G1	6.92	3.6	164
S	8.30	3.1	102
G2	1.77	2.2	102
Cell Cycle	17.1	2.1	102

Table : On Michael Olma set with the H2B channel

Outline

Cancer research

1. Cancer is a disease of the genome.
   In the early 2000, first sequencing of the human genome.
2. Cancer is a very heterogeneous disease. 2 breast cancers can be totally different on a molecular basis. This disease is highly variable in many aspects, we therefore saw the emergence of personalized medicine.
3. The clinical motivation for cancer research are: profiling each type of tumor, in terms of molecular subtype and prognostic in order features. On the long run this will lead to a better understanding, diagnostic, prognostic, ...

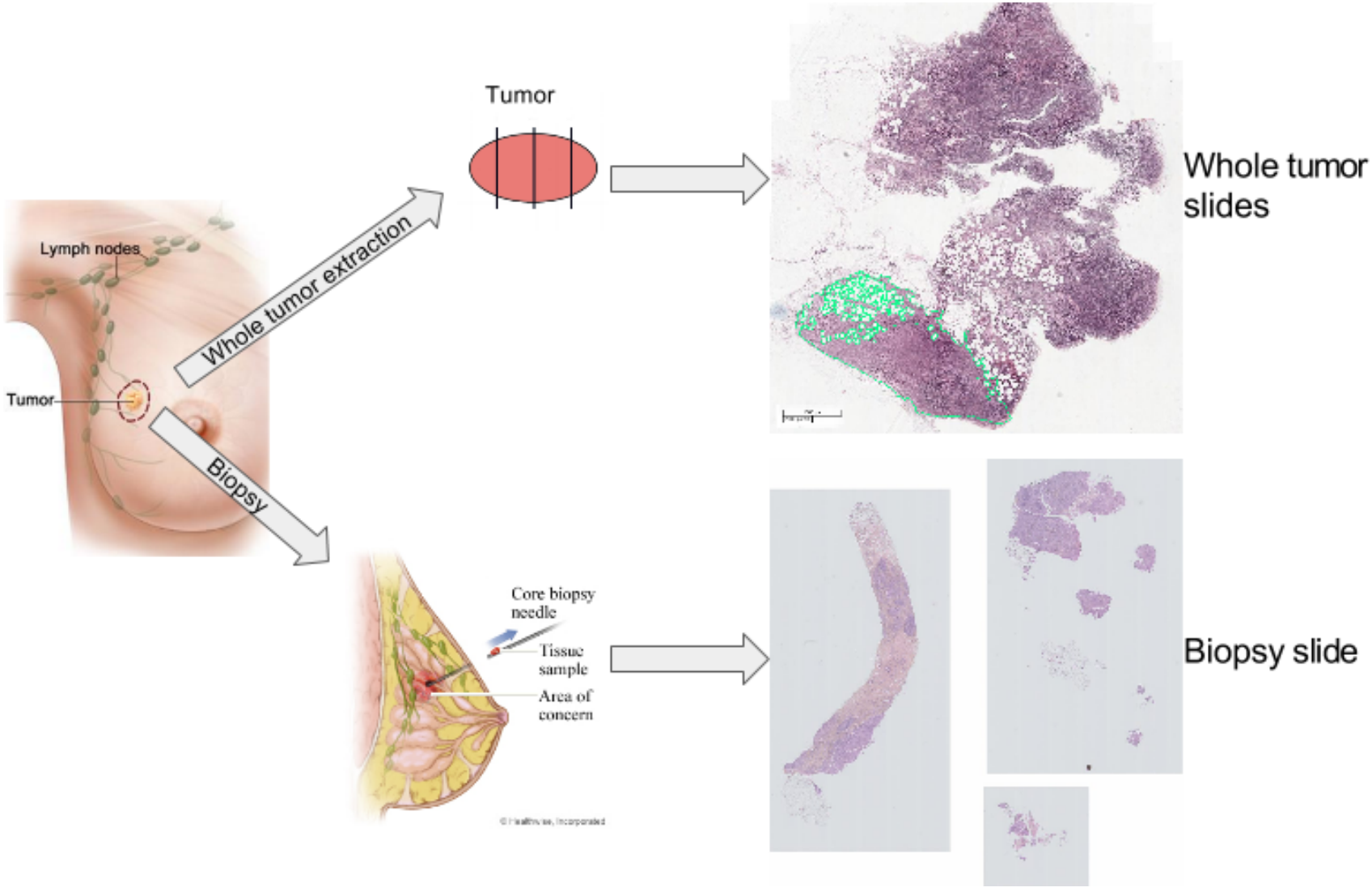
Why image based features in cancer research

- In clinical practice, one looks at the available image data to give a treatment.
- Differently, in clinical research, one studies the images to infer properties.
- Finding correspondence/correlations between genes and phenotypic information.
- Highly relevant data can be found.
  - We can cross protein localization microscopy with loss-of-function experiment.
  - You access response to disease information, lymphocyte responses to tumor expansion. Different type of information, spatial repartition, neighbouring information, density, ..
  - Correcting noisy molecular data, mutation data will be mixture of healthy cells, cancerous cells, lymphocytes,...

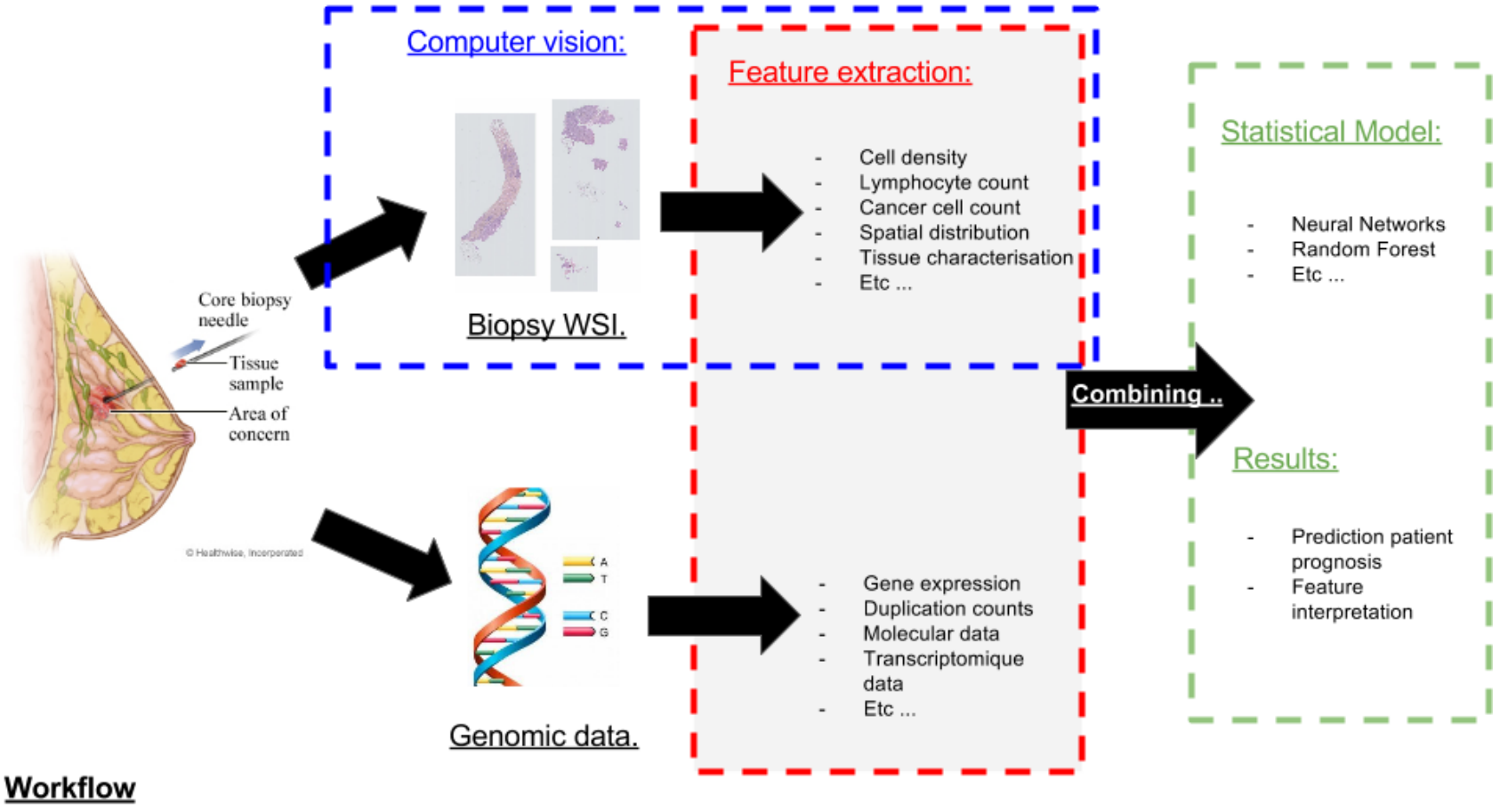
Outline

Histopathology: microscopic examination of tissue

Different procedure of histopathology in breast cancer.



PhD pipeline



Main focus

Extracting relevant features at the cellular and tissue level.

Cellular level:

- Segmentation at the cellular level.
- Classification of nuclei type: Lymphocyte, cancerous cell, normal cell, ...
- Classification of nuclei state: Cell undergoing mitotic event, dead cell, ...

Tissue level:

- Segmentation at the tissue level.
- Classification into regions: tumor, stromal and necrotic. Information at the cellular level could be used to help/correct prediction.

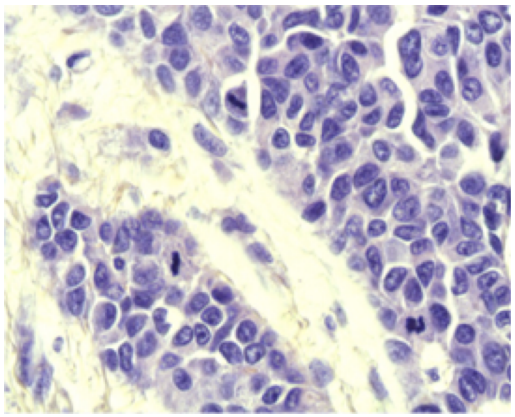
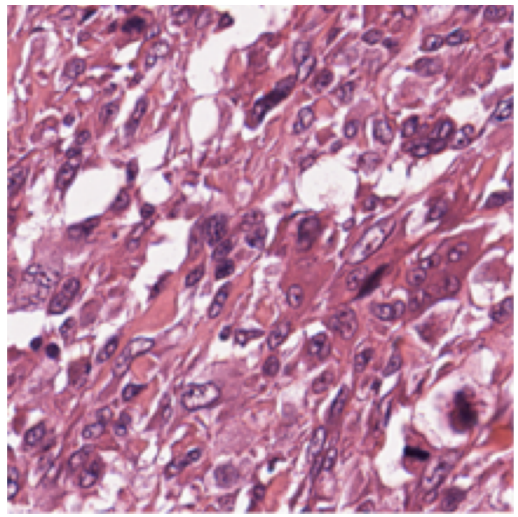
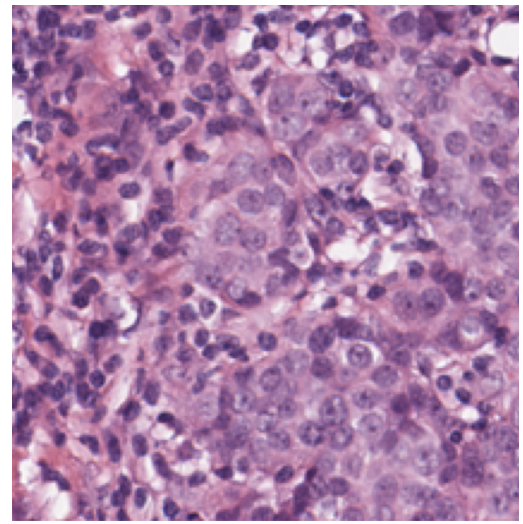
Main issue: As we wish this to be easily reproducible, the key problem is the segmentation.

Application to two datasets

1. 208 slides from an unpublished study on breast cancer, a special type of very aggressive breast cancer
   Focus: treatment response to chemotherapy.
2. 198 slides from a recently published study on bladder cancer[1].
   Focus: Correlation between histopathology features with the molecularly defined subgroups.

[1] Anne Biton et al. "Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes". In: *Cell reports* 9.4 (2014), pp. 1235–1245.

Difficulties

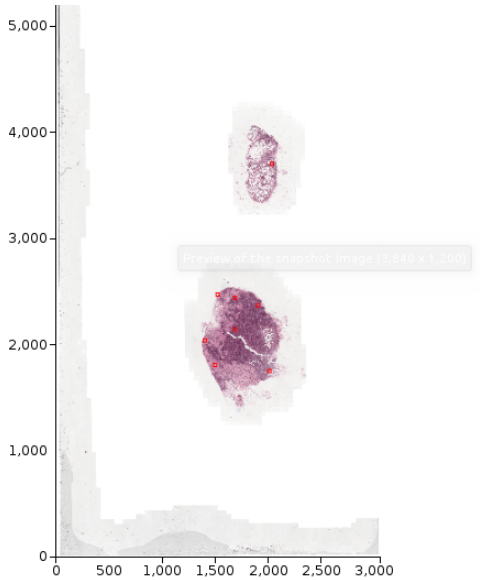




Tissue sections with standard Haematoxylin and Eosin staining

- High variability in staining.
- High variability in tissue objects.
- Difficult segmentation task.

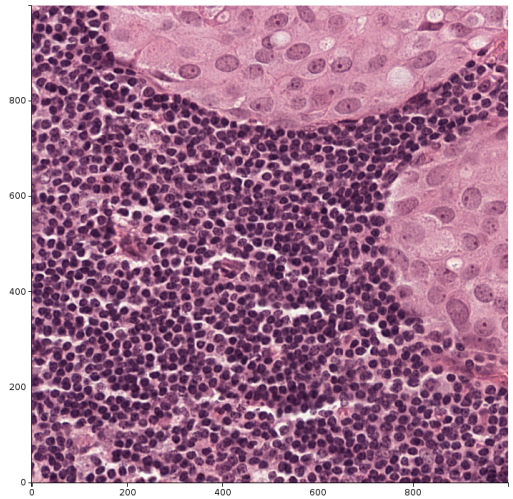
Large image files: tiled tiff format

One individual slide can be up to $50GB$, uncompressed.



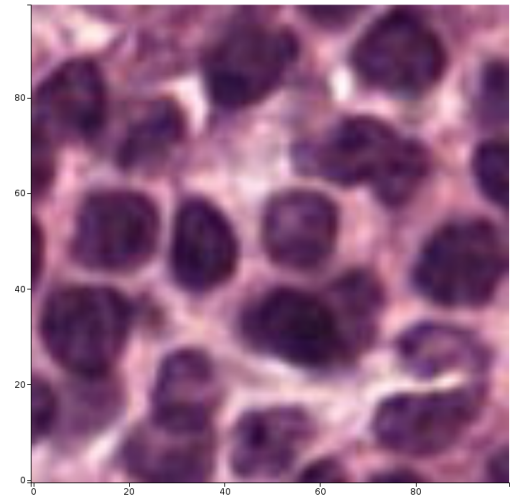
Tumor 31

7000 x 3500, resolution 6



Sub-image of Tumor 31, highest resolution

1000 x 1000, resolution 0



Sub-image of Tumor 31, highest resolution (zoom)

100 x 100, resolution 0

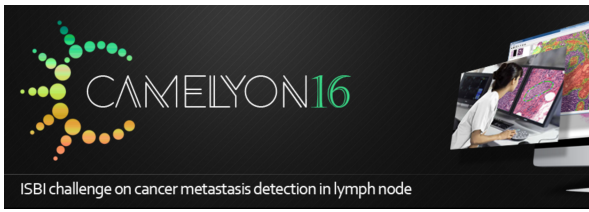
Outline

Camelyon2016

In many machine learning tasks, annotated data is very expensive. They often require experts and these task are long, tedious and prone to manual errors. Camelyon2016 provides:

- Over $500GB$ of histopathology slides. (400 slides)
- 270 fully annotated slides, for metastasis detection.
- First experience with:
  - Histopathology data
  - Computer vision
  - Dealing with huge databases (clusters of computers)

Work in collaboration with V. Machairas, E. Decenciere and T. Walter.



Official logo

Pixel based classifier

Finding metastasis regions:

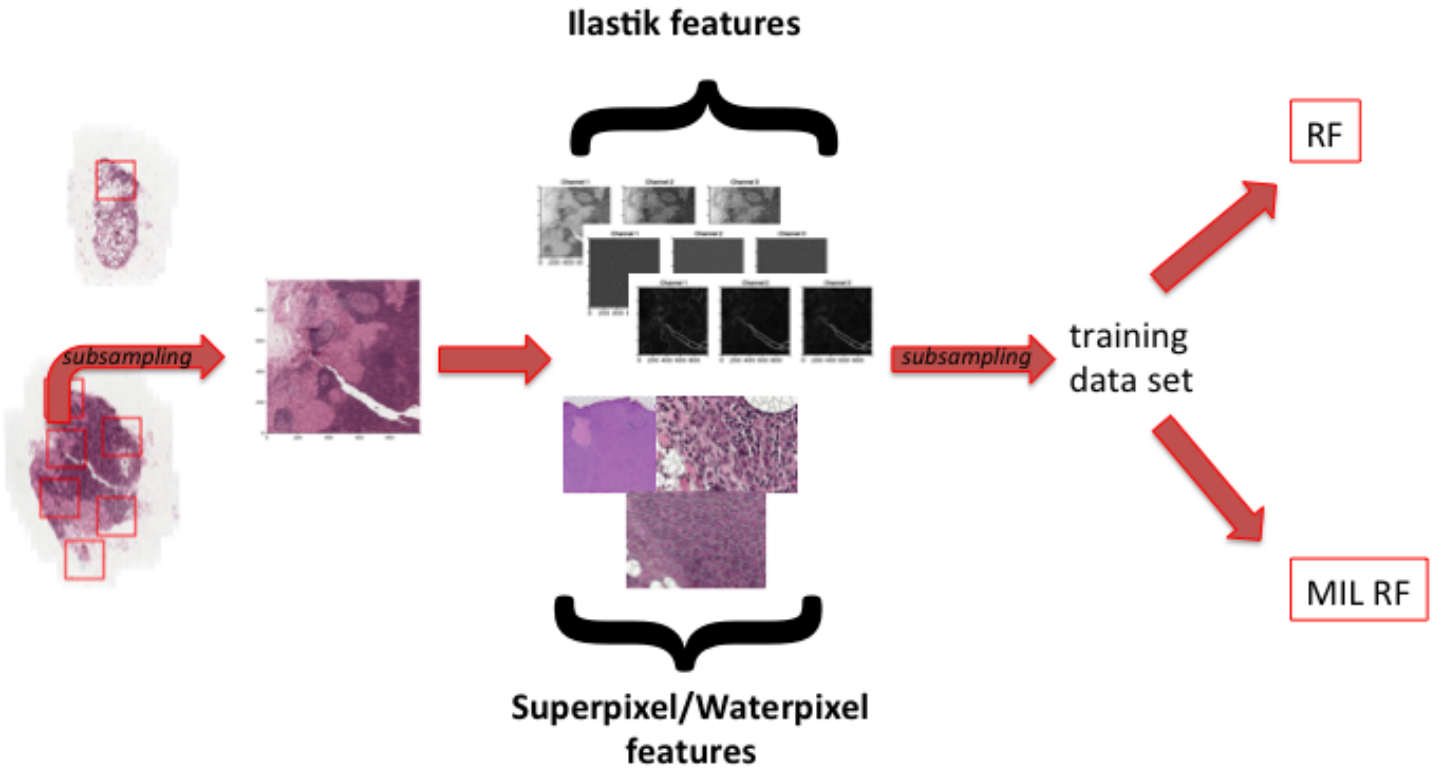
Each pixel, will be considered metastasis if it belongs to a metastasis region.

Model: Modified random forest, each trees see a unique part of the data set.

Cross-validation/evaluation:

One slide out scheme.

Work pipeline



Camelyon2016 pipeline

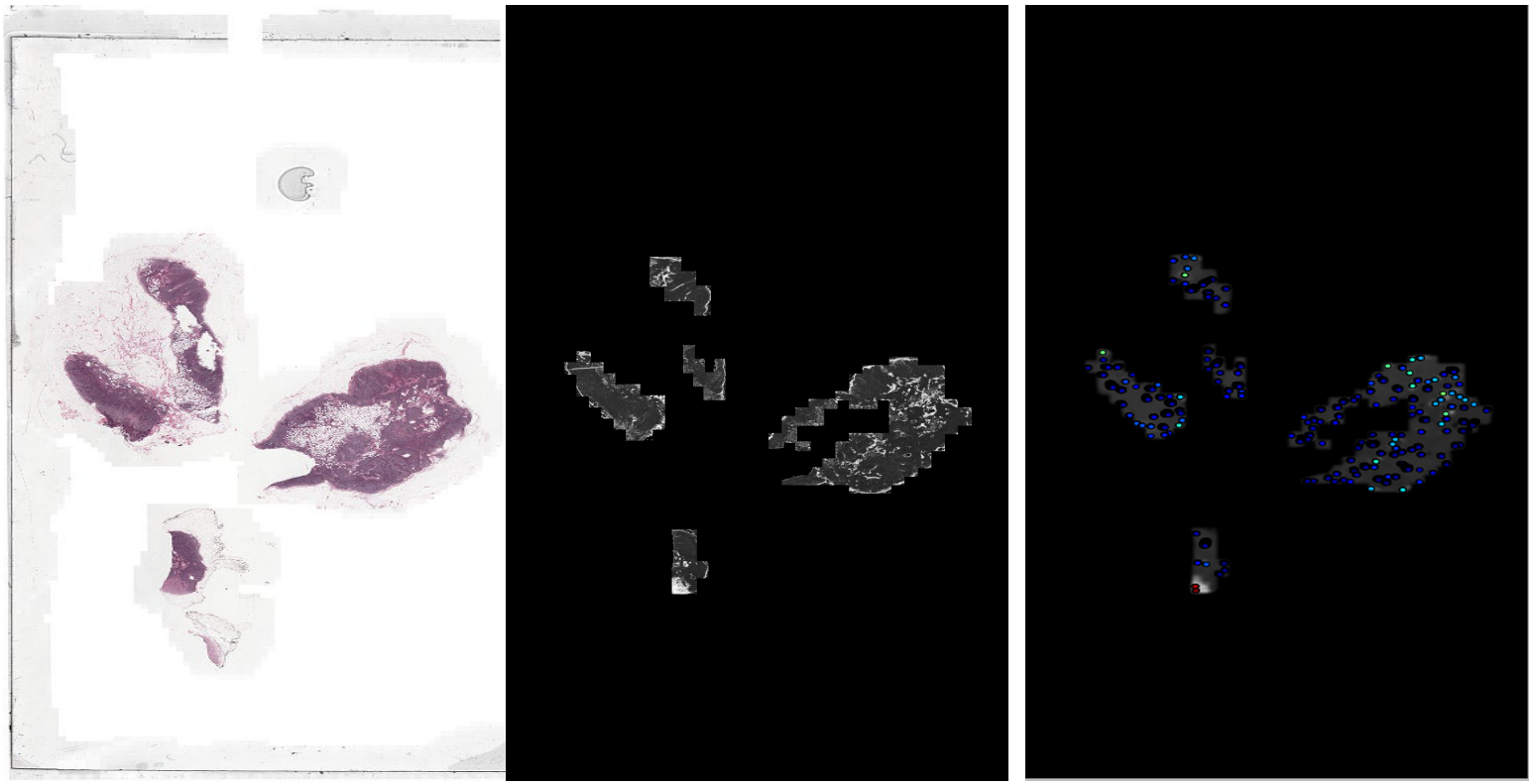
Ilastik

Ilastik : software for interactive image classification, segmentation and analysis.
Using features from ilastik (implemented in vigra, c++ library)

- Color/Intensity:
  - Gaussian Smoothing
- Edge:
  - Laplacian of Gaussian
  - Difference of Gaussians
  - Gaussian Gradiant Magnitude
- Texture:
  - Structure Tensor Eigenvalues
  - Hessian of Gaussian Eigenvalues

Ilastik

Pixel based classifier: outputs



(a) RGB raw data. (b) Probability map. (c) Prediction points.

For figure (c), blue is equal to a low confidence score whereas red means a high confidence.

Evaluation whole slide image, test slide number 2.

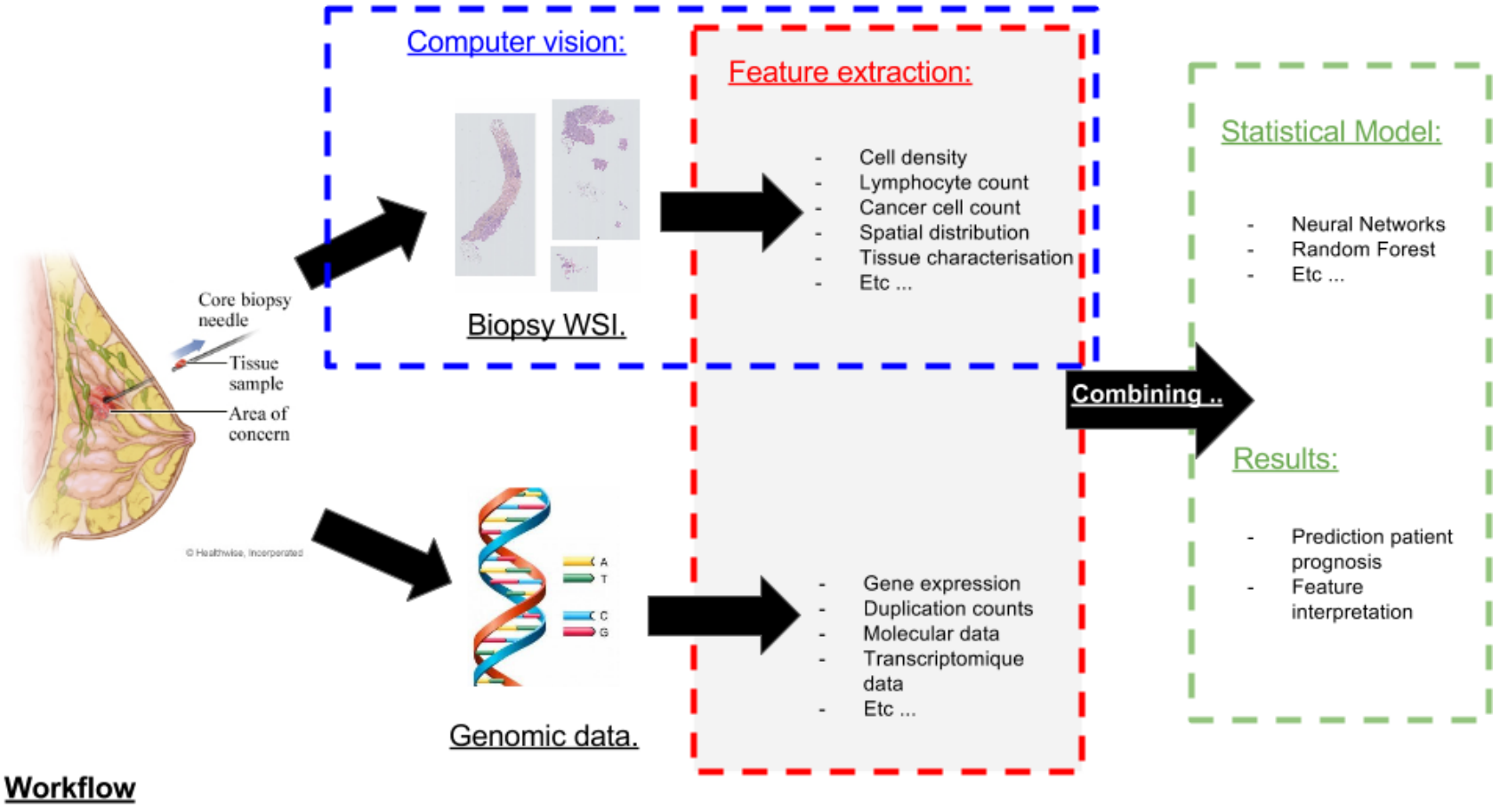
Results and perspective

- We had an AUC score of $0,63$.
- Our features are based ilastik/waterpixels at different scales.
- Unsufficient to predict the higlhly variable metastasis tissue.
- Methods based on deep architecture achieved good performances.
- Learning which features are relevant is the key point.
- Next step: segmentation based on deep architecture, fully convolutionnal networks[2].

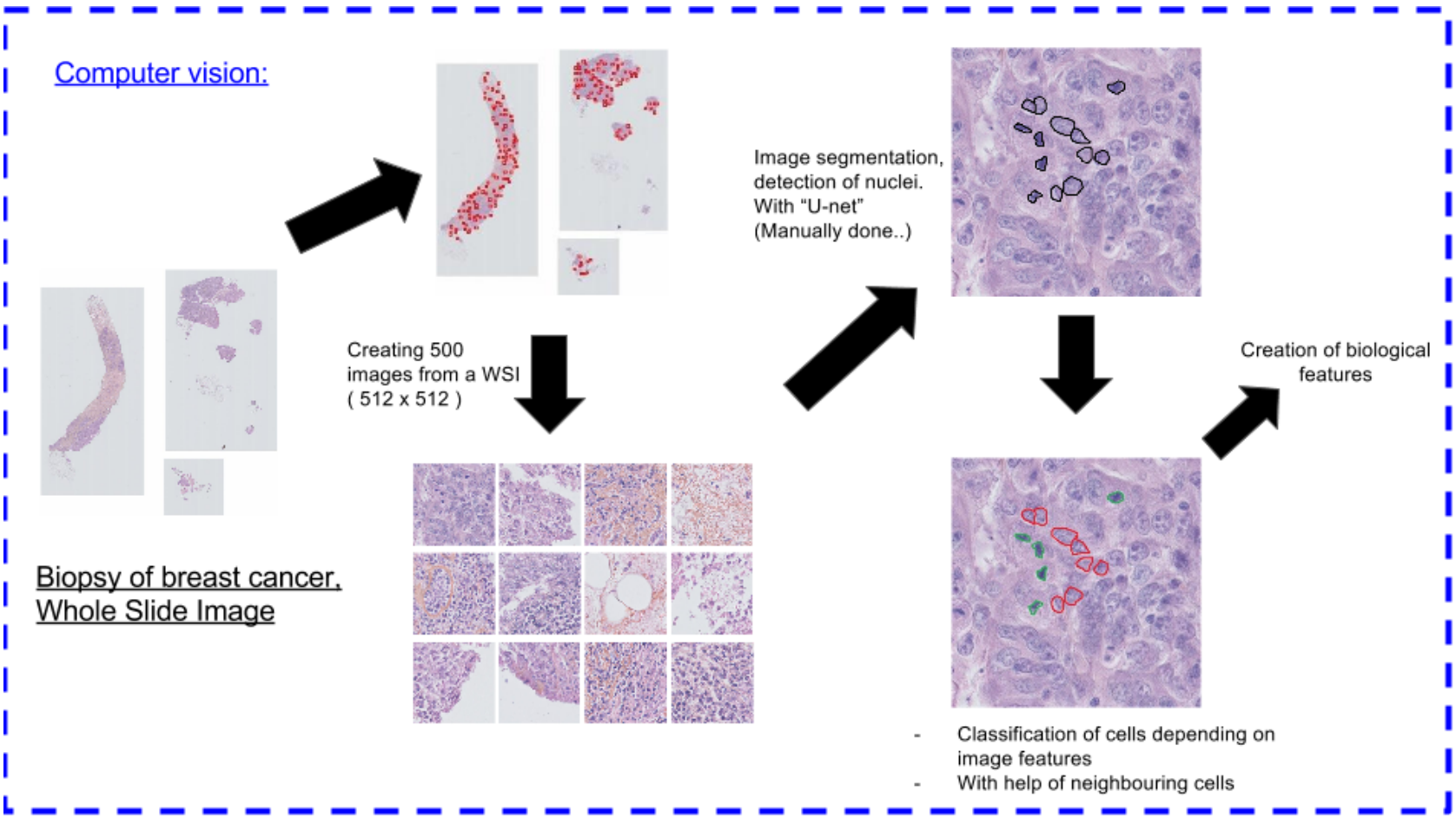
[2] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 3431–3440.

Outline

PhD pipeline

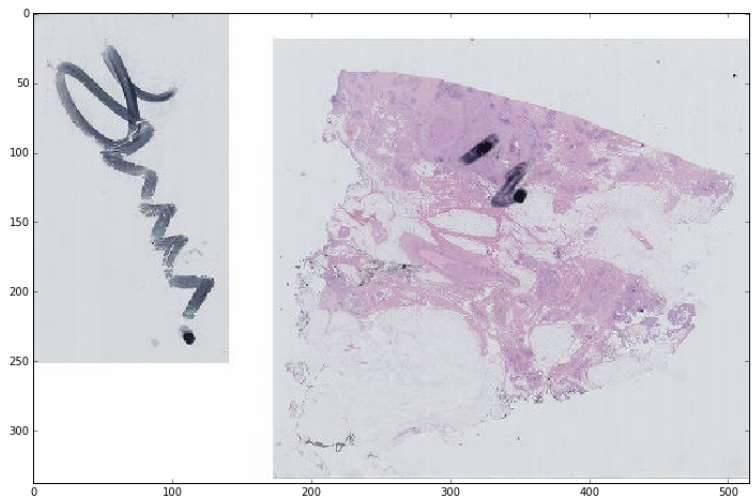


Computer Vision

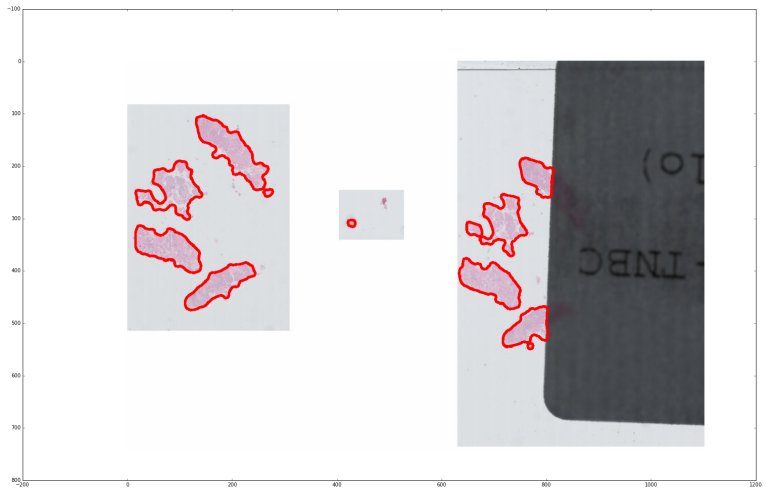


Data

Very similar data as Camelyon2016, but messier. We had to better segment the tissue area.



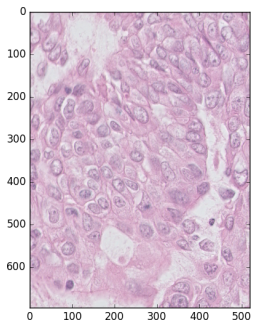
Whole slide tumor

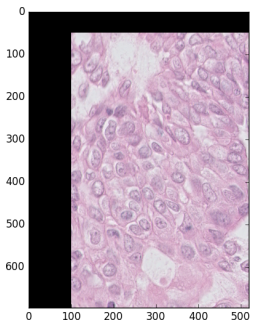


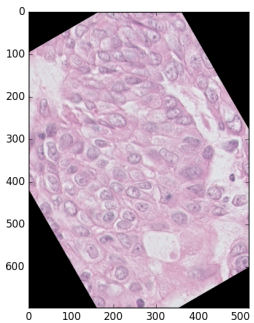
Biopsy

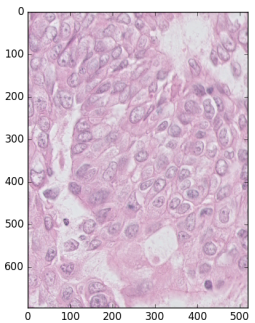
Annotations

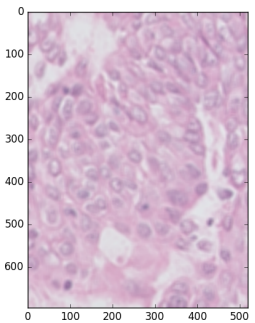
- Manual annotations at the cell/tissue level.
- Issue very little ground truth while deep networks need a lot of data? Data augmentation!

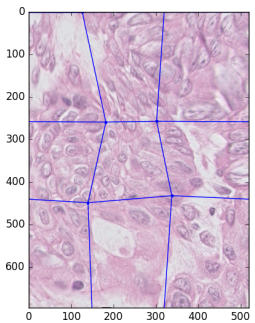

Original


Translated


Rotated


Flipped


Blurred


Elastic distortion

Data augmentation can make the network learn the proper invariances![3].

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*. Springer, 2015, pp. 234–241.

Outline

Jupyter notebook and Github

`https://github.com/PeterJackNaylor?tab=repositories`

GitHub, Inc. [US] https://github.com/PeterJackNaylor/PhD_Fabien/blob/master/AssociatedNotebooks/DataAugmentation.ipynb

Data augmentation

As the segmentation data is scarse, a trick to have more avaibable annotated data is to use data augmentation. Data augmentation is also a way of teaching certain invariance to your machine learning algorithm.

In particular, for biomedical data it is interesting to have these invariance:

- Translation
- rotation
- Miror
- Out of focus or bluriness (due to the scanner for instance)
- Elastic deformation

Loading the toy data

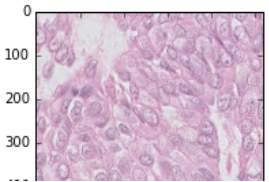
In [5]:
```
%matplotlib inline
### Useful plotting function
def plot_comparison(original, modified, modification):

    fig, (ax1, ax2) = plt.subplots(ncols=2, figsize=(16,16), sharex=True,
                                   sharey=True)
    ax1.imshow(original, cmap=plt.cm.gray)
    ax1.set_title('original')
    ax1.axis('off')
    ax1.set_adjustable('box-forced')
    ax2.imshow(modified, cmap=plt.cm.gray)
    ax2.set_title(modification)
    ax2.axis('off')
    ax2.set_adjustable('box-forced')
```

In [6]:
```
from skimage.io import imread
import matplotlib.pylab as plt


sample = imread('/home/naylor/Bureau/209.png')
## It has fourth (useless) component
sample = sample[:,:,0:3]

plt.imshow(sample)
plt.show()
```



The end! :-)

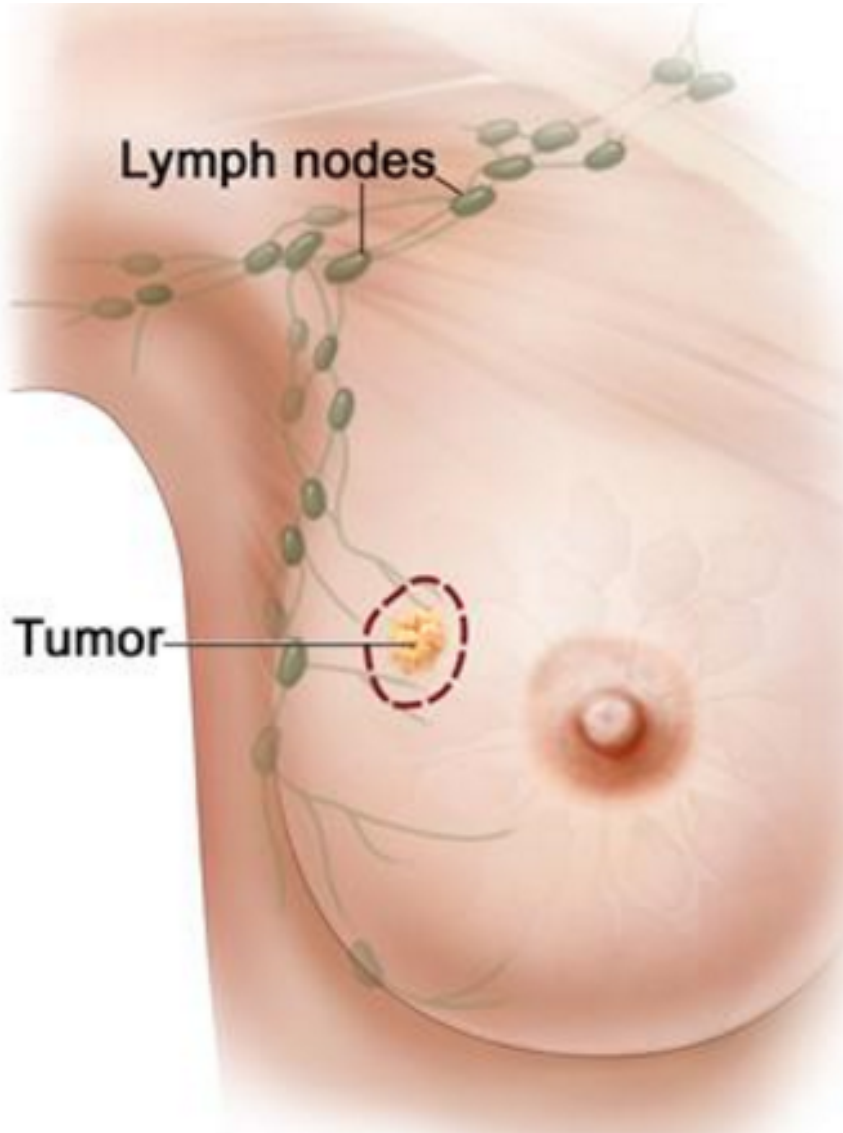
Camelyon16

Goal:

- Detection of micro- and macro-metastases in lymph node digitized images.
- Automated detection of metatases in hematoxylin and eosin (H&E) stained whole-slide images of lymph node sections.

Motivation:

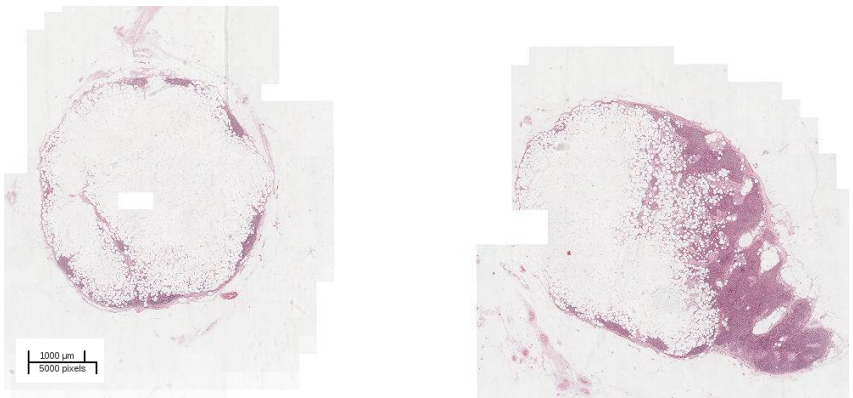
- Lymph node metastases occur in most cancer types.
- Lymph nodes in the underarm are the first place cancer is likely to spread.
- The prognosis is poorer when cancer has spread there.



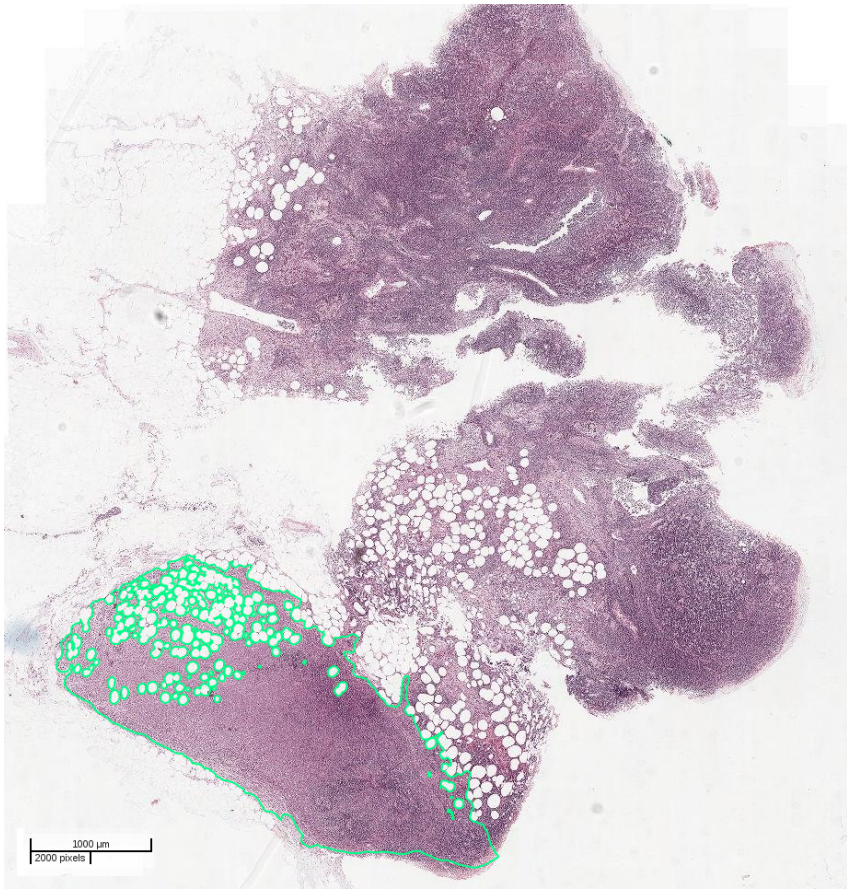
Camelyon16

Goal:

- Detection of micro- and macro-metastases in lymph node digitized images.
- Automated detection of metastases in hematoxylin and eosin (H&E) stained whole-slide images of lymph node sections.



Normal 38



Tumor 34

Metastases detected in green

Evaluation

Two evaluation metrics and two leader boards.

Slides based:

- Binary classification of whether or not a slide contains metastases.
- Evaluating with the area under the ROC curve.

Region based:

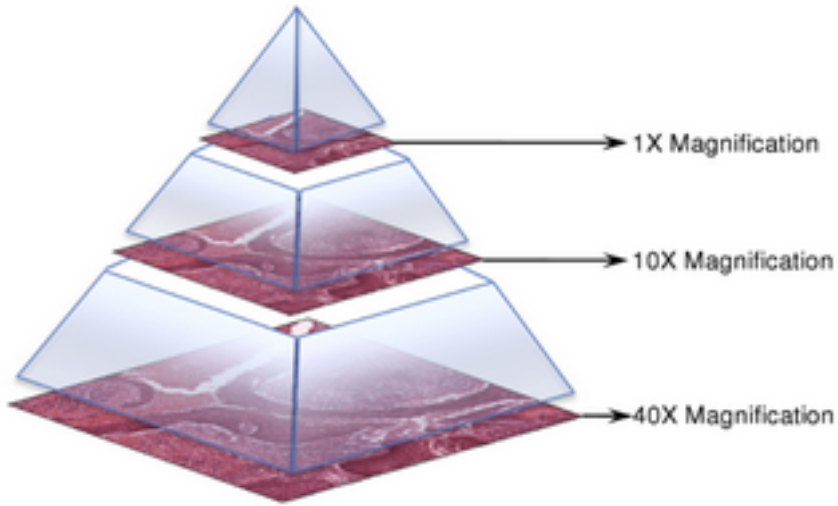
- Correctly detecting metastases within slides.
- Evaluation with the FROC curve (free-response receiver operating characteristic)

The data set

Data set provided:

- 160 Normal slides.
- 110 Tumor slides.

  - Huge images with very high precisions, using c++ library openslide.
  - One image compressed with JPEG2000: $\sim$ 2/3Gb.
  - Uncompressed at approx. 10Gb.
  - Highest resolution : $\sim 96256 \times 218624$.
  - Lowest resolution : $\sim 188 \times 427$.



Pyramid data structure

Between 8 and 10 different resolutions

Images

Very large number of samples:

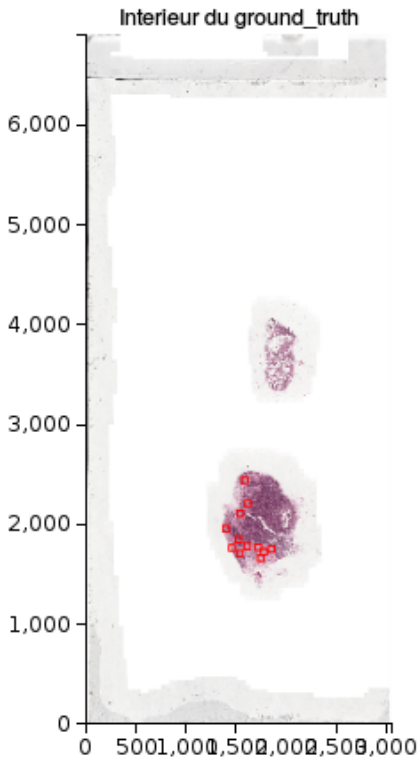
- In pixel classification, each pixel is an instance, $n \gg p$.
- Appropriate subsampling methods. In particular here, first subsampling for the "sub-images". Second subsampling given a particular "sub-image".

Subsampling 1

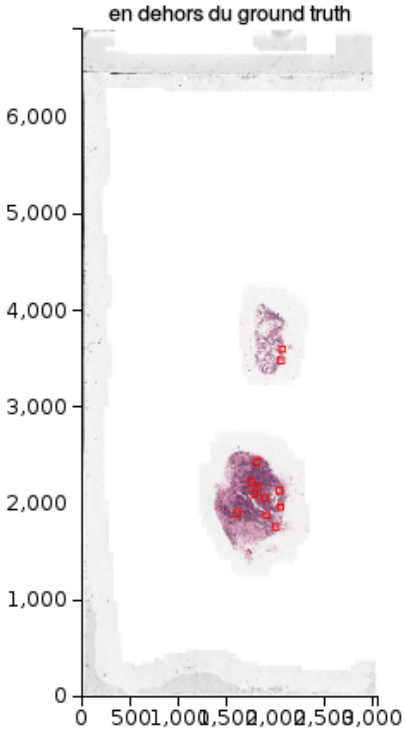
Subsampling 1 (region of interest detection over slides) Trying to find interesting part of the image.
We can divide sub-images in 4 groups.

- Only metastasis tissue.
- Only normal tissue.
- Centered on the boarder metastasis tissue/normal tissue.
- Centered on the boarder tissue/background.

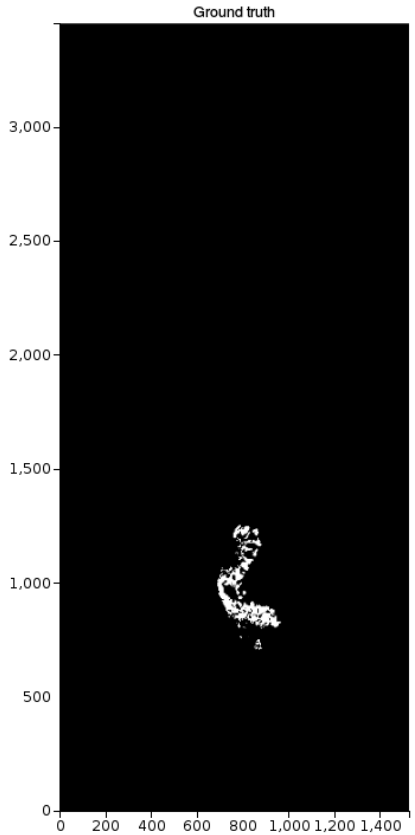
Subsampling



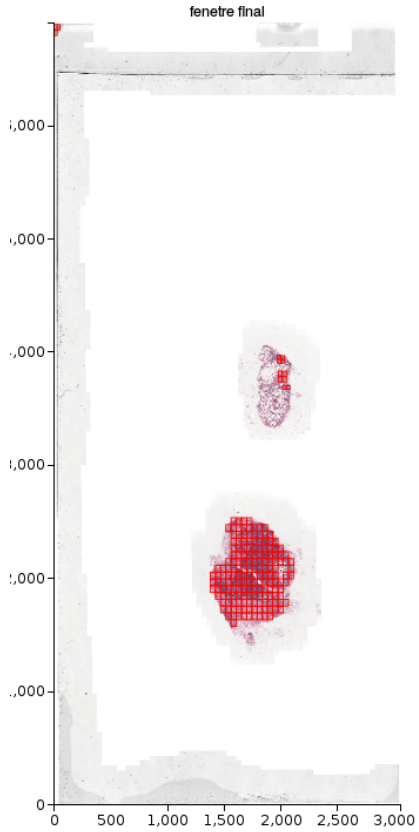
Only metastasis regions



Only metastasis-free regions

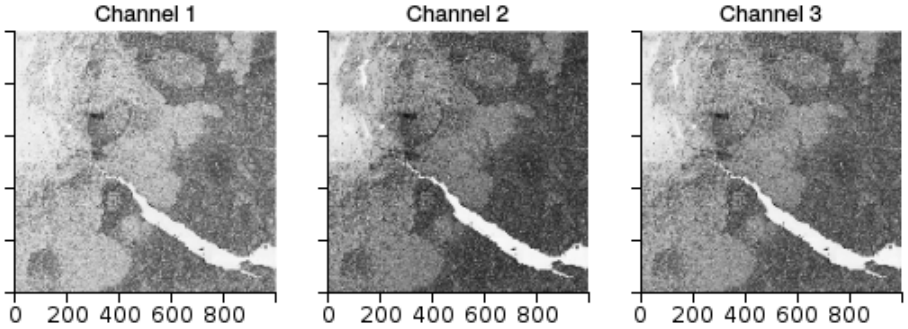


Ground truth



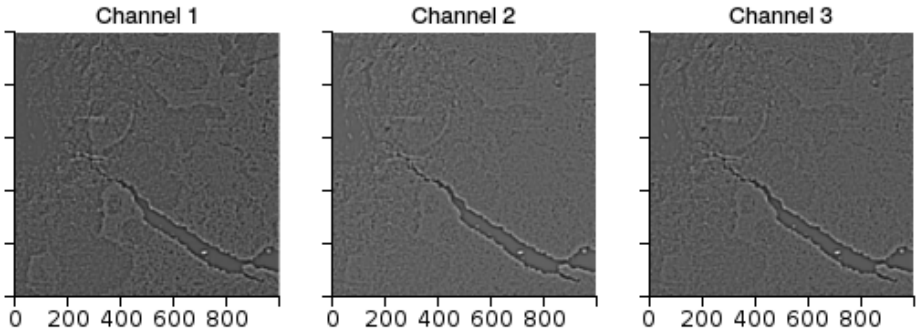
Grid partition

Ilastik - examples



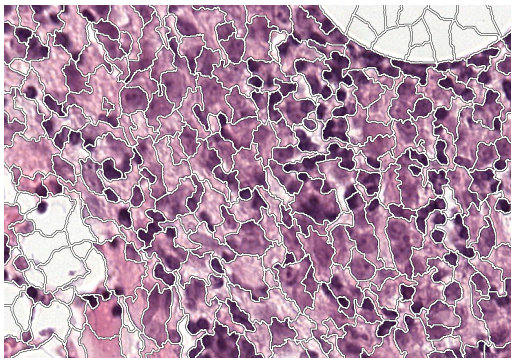
Gaussian smoothing, $\sigma = 0.7$

Ilastik

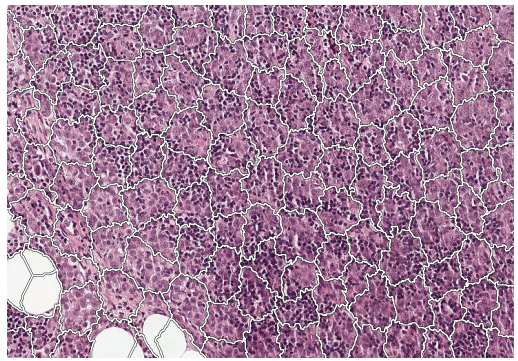


Laplacian of Gaussian, $\sigma = 5$

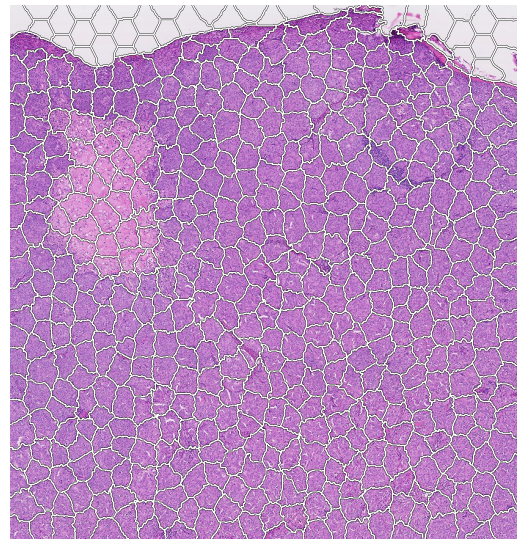
Waterpixels - examples



Waterpixels at resolution 0



Waterpixels at resolution 2



Waterpixels at resolution 4

Subsampling 2

Pixel classification.
Still very large data-set.
260 slides $\rightarrow$ 100 sub-images per slides $\rightarrow$ 1 000 000 pixels per sub-images.
Need for a second resampling.
Resampling randomly? smartly, via unsupervised methods? K means algorithm?

Pixel based classifier

Finding metastasis regions:
Each pixel, will be considered metastasis if it belongs to a metastasis region.
Model: random forest.
Detecting metastasis patients: Each pixel is or not a metastasis but it also belongs to bigger group. This larger group, the slide, is annotated.
Model: Multiple instance learning random forest.
Cross-validation/evaluation:
One slide out scheme.

Multiple instance learning

Notations:

- Pixels are a pair $(x_i, y_i) \in \mathbb{R}^d \times \{-1, +1\}$.
- Slides are bags of pixels: $B_I = \{x_i, i \in I\}$, and we have $Y_I = 1$ if there is at least one x_i in B_I that is positive, otherwise $Y_I = -1$.
- Constraint to add to an optimization problems:

$$\sum \frac{y_i + 1}{2} \geqslant 1, \forall I \ s.t \ Y_I = 1 \text{ and } y_i = -1, \forall I \ s.t \ Y_I = -1$$

Implementation : MISVM, Multiple-Instance Support Vector Machines by Gary Doran. Python package.
Paper: *Support Vector Machines for Multiple Instance learning*, S. Andrews, I. Tsochantaridis and T. Hofmann.

Superpixels/Waterpixels

Superpixels: Regions resulting from a low-level segmentation.
They have these proprieties: homogeneity, connected partitions, adherence to object boundaries, regularity.

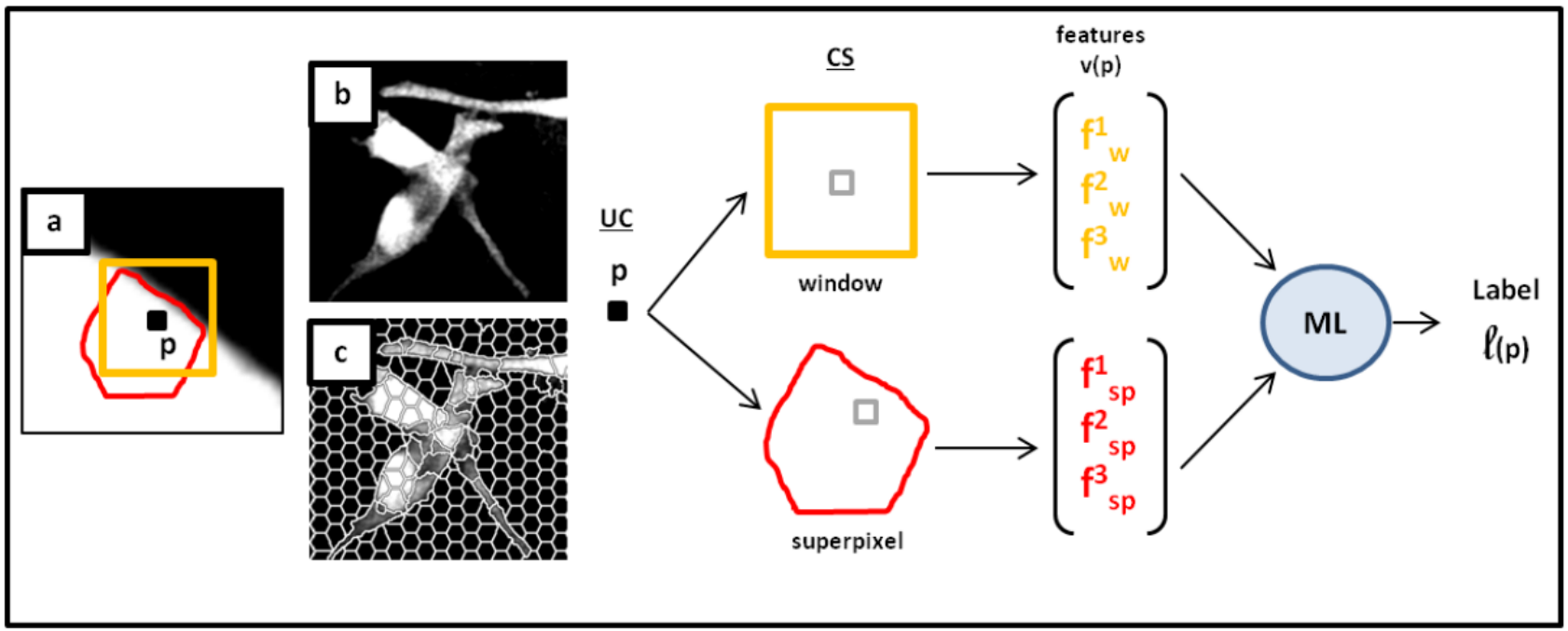


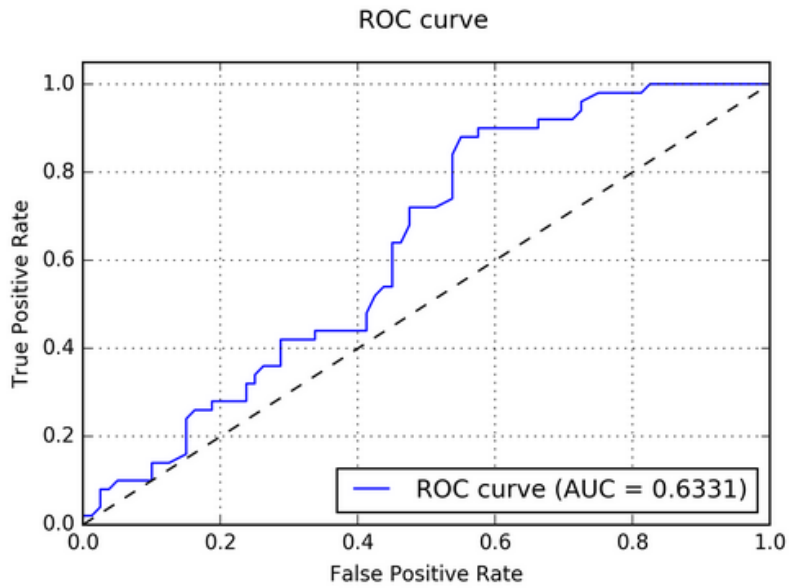
Illustration of how superpixels are used

~~For more details~~[4].

[4] Vaïa Machairas et al. "Waterpixels". In: *Image Processing, IEEE Transactions on* 24.11 (2015), pp. 3707–3716.

Challenge results

We did a score of 22nd out of 23.
Only three teams performed non-deep convolutionnal network. We were 2nd among these 3 teams.



ROC curve associated to the slide base evaluation