Predicción de penetración del adenocarcinoma en el colón mediante AI

ikerbasque
Basque Foundation for Science

MAX-PLANCK-GESELLSCHAFT

Universidad del País Vasco
Euskal Herriko Unibertsitatea

HPC Admintech 2022:
Palma de Mallorca
11, **12**, 13 y 14 de mayo
Workshop de HPC para la Ciencia

Marcos J. Araúzo Bravo
Julen Bohoyo Bengoetxea

Jose J. Rodriguez Anda

biodonostia
osasun ikerketa institutua
instituto de investigación sanitaria

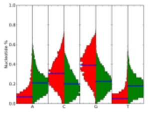
bioaraba
osasun ikerketa institutua
instituto de investigación sanitaria

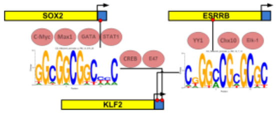
Big data for biomedicis: Some algorithms

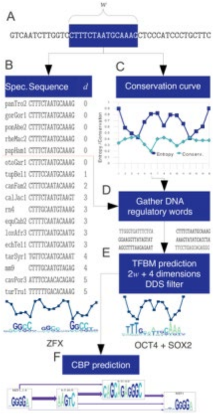
Prediction of DNA motifs

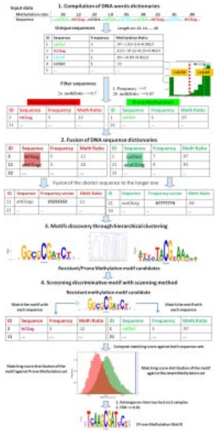
TF binding motifs

DNA methylation motifs






A
B
C
Spec. Sequence
Conservation curve
D
Gather DNA regulatory words
E
TFBM prediction 2w + 4 dimensions DDS filter
ZFX
OCT4 + SOX2
F
CBP prediction

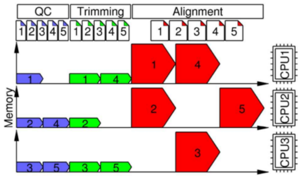

1. Compilation of DNA words dictionaries
2. Fusion of DNA sequence dictionaries
3. Motifs discovery through hierarchical clustering
4. Screening discriminative motif with scanning method

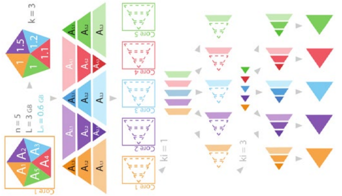
Müller-Molina et al, ***PLoS ONE***, 2012

Luu et al, ***Genome Research***, 2013

Luu et al, ***Bioinformatics***, 2016

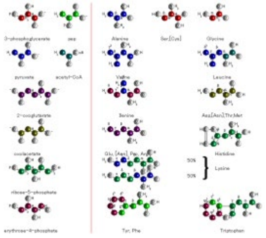
Ascension & Araúzo-Bravo, ***IEEE/ACM Tran. Com. Bio.***, 2020

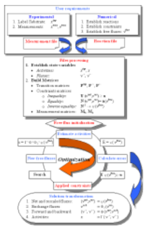


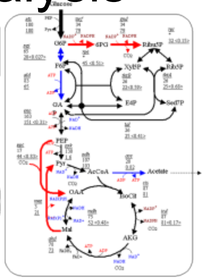


Metabolic engineering

Metabolic flux analysis







Araúzo-Bravo and Kazuyuki, ***Journal of Biotechnology***, 2003

Zaid et al, ***FEMS Microbiology Letters***, 2004

Zaid et al, ***Applied Microbiology and Biotecnology Letters***, 2004

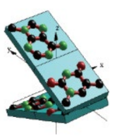
Peng et al, ***FEMS Microbiology Letters***, 2004

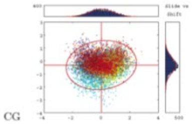
Sarkar et al, ***Archives. of Microbiology***, 2008

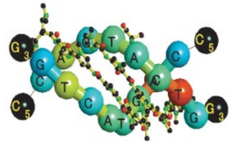
Structural biology

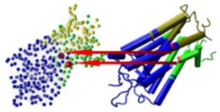
DNA proteins and drugs interactions

Protein communications









Araúzo-Bravo et al, ***Journal American Chemical Society***, 2005

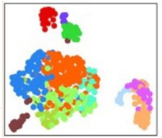
Ahmad et al, ***Nucleic Acid Research***, 2006

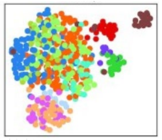
Del Sol et al, ***Genome Biology***, 2007

Araúzo-Bravo et al, ***Nucleic Acid Research***, 2008

Single cell omics

Feature selection





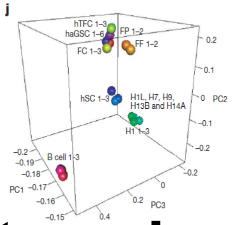
Ascension et al, ***Gigascience***, 2022

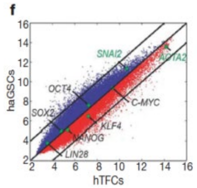
Big data for biomedics: Some results

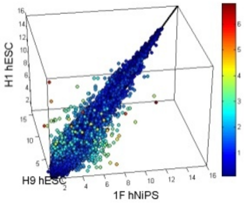
Kim et al, ***Nature***, 2008
Kim et al, ***Cell***, 2009
Kim et al, ***Nature***, 2009
Ko et al, ***Nature***, 2010
Han et al, ***Cell***, 2010
Han et al, ***Nature Cell Biology***, 2011
- Knochbloch et al, ***Nature***, 2012

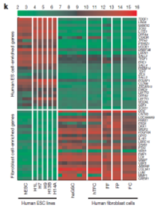
Moore et al, ***Science***, 2015
Rao et al, ***Cell Stem Cell***, 2016
Song et al, ***Cell Stem Cell***, 2016

Transcriptomics

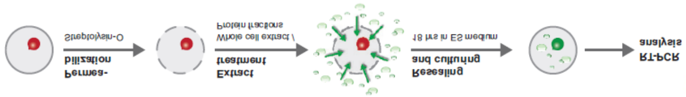


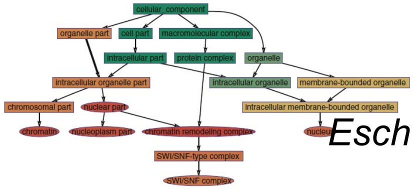






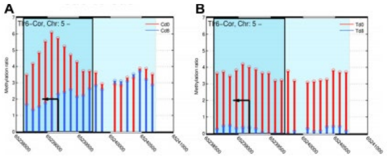
Proteomics

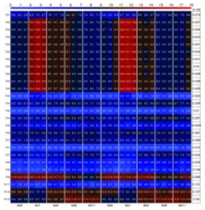




Singhal et al, ***Cell***, 2010
Esch et al, ***Nature Cell Biology***, 2012

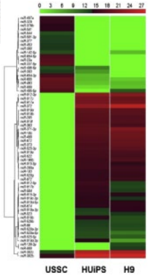
Methylomics





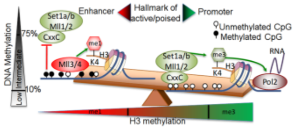
Santourlidis et al, ***Stem Cell Res***., 2011
Hargus et al, ***Cell Reports***, 2014
Al-Quraishy, ***Parasitology Research***, 2014
Dhikl et al, ***J. Steroid Bioch. Mol. Biology***, 2015
Dorn et al, ***Haematologica***, 2015
Luu et al, ***Bioinformatics***, 2016

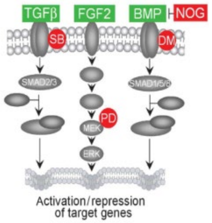
Micro RNA expression



Zaehres et al, ***Exp. Hematology***, 2010

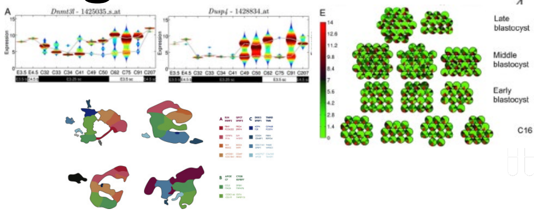
CHiP-Seq

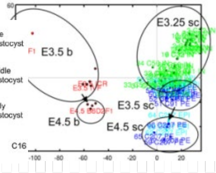




Greber et al, ***EMBO***, 2011

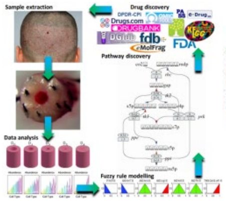
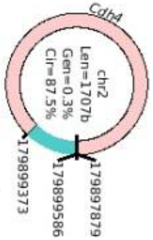
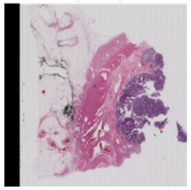
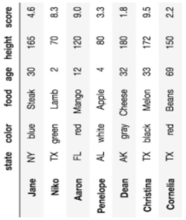
Single-cell omics





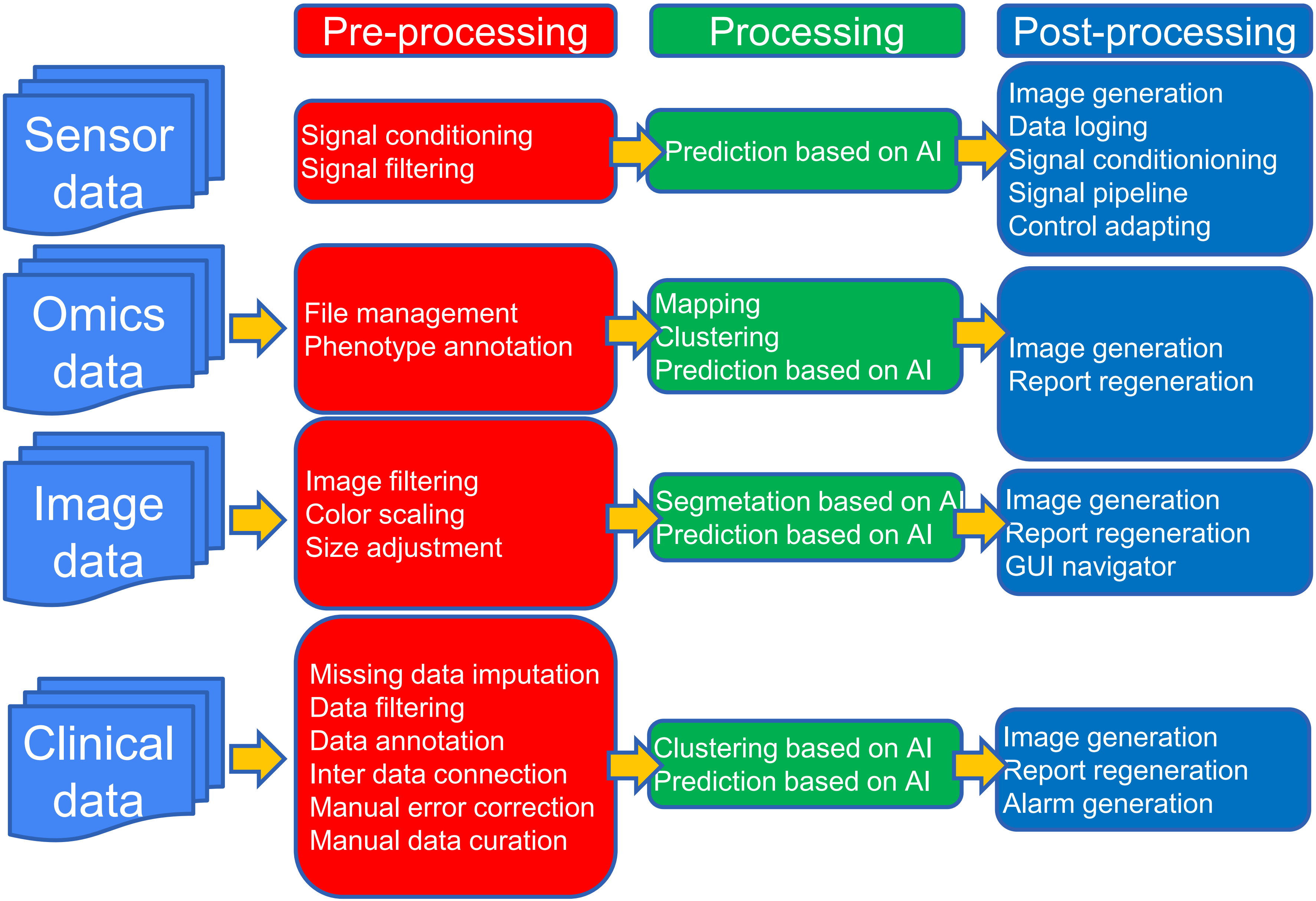
Grinberg et al, ***PNAS***, 2013
Ohnishi et al, ***Nature Cell Biology***, 2014
Gerovska & Araúzo-Bravo, ***Mol. Human Reproduction***, 2016
Ascension et al, ***J. Investigation Dermatology***, 2020

Some of our past and ongoing AI projects

	Examples		Technology	Project	Funding
Sensor data			Fuzzy Logic Artifical Neural Net	PSYCHO MONNET	European Commission
Omics data	 Sample extraction Drug discovery Drugs.com DRUGBANK fdb eMolFrag FDA KEGG e-Drug Pathway discovery Data analysis Fuzzy rule modelling	 Cdh4 chr2 Len=1707b Gen=0.3% Cir=87.5% 179893373 179899586 179897879	Fuzzy Logic Random forest Artifical Neural Net	4D Healing Circular Vision	European Commission
Image data			Deep Learning	PreCCol	EUSKO JAURLARITZA GOBIERNO VASCO
Clinical data			Random forest Genetic algorithms Deep Learning	STRATOS	GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

	state	color	food	age	height	score
Jane	NY	blue	Steak	30	165	4.6
Niko	TX	green	Lamb	2	70	8.3
Aaron	FL	red	Mango	12	120	9.0
Penelope	AL	white	Apple	4	80	3.3
Dean	AK	gray	Cheese	32	180	1.8
Christina	TX	black	Melon	33	172	9.5
Cornelia	TX	red	Beans	69	150	2.2

Common steps in AI data analysis projects



Pre-processing

Processing

Post-processing

Sensor data

Signal conditioning
Signal filtering

Prediction based on AI

Image generation
Data loging
Signal conditioning
Signal pipeline
Control adapting

Omics data

File management
Phenotype annotation

Mapping
Clustering
Prediction based on AI

Image generation
Report regeneration

Image data

Image filtering
Color scaling
Size adjustment

Segmetation based on AI
Prediction based on AI

Image generation
Report regeneration
GUI navigator

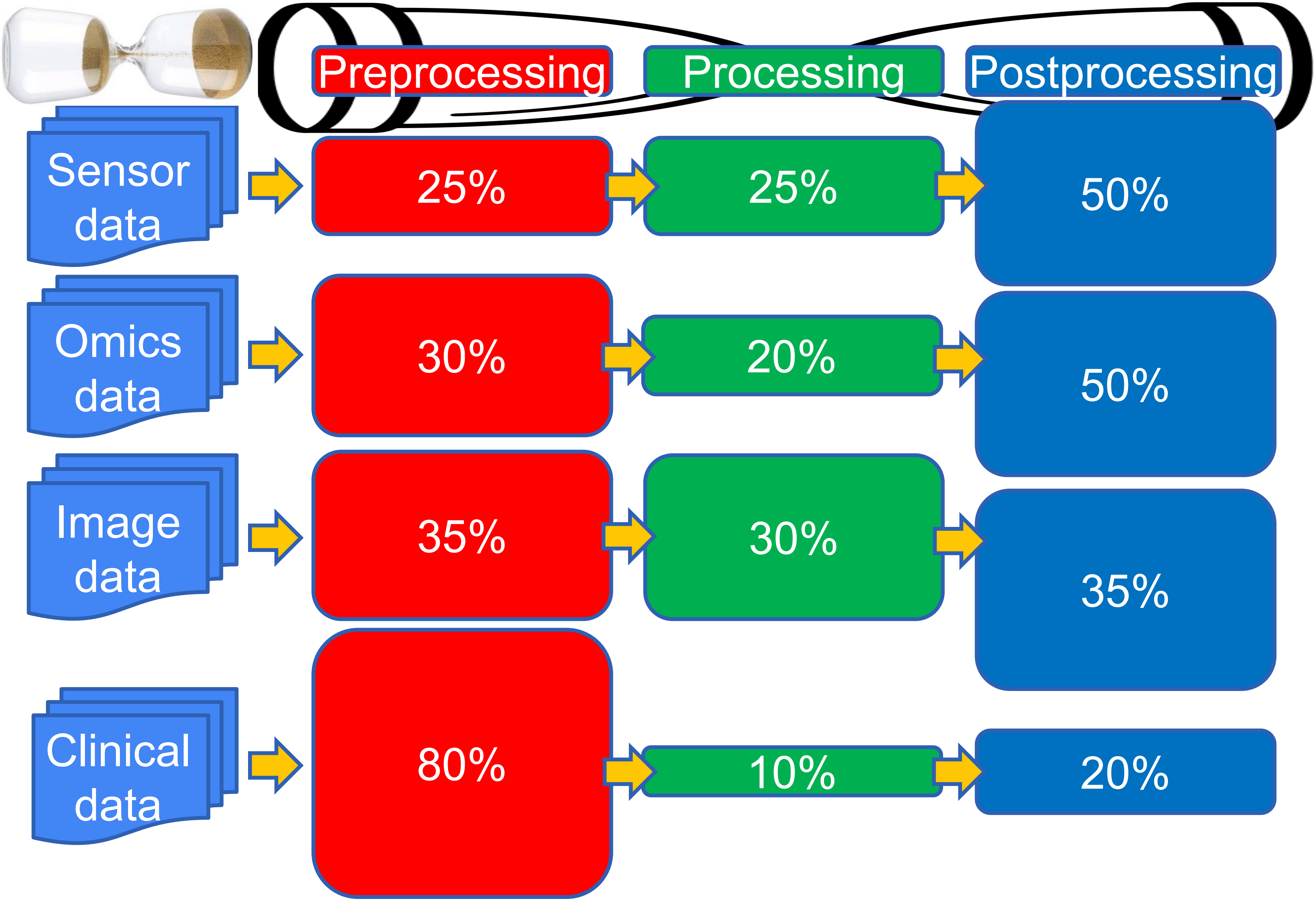
Clinical data

Missing data imputation
Data filtering
Data annotation
Inter data connection
Manual error correction
Manual data curation

Clustering based on AI
Prediction based on AI

Image generation
Report regeneration
Alarm generation

Typical timing of AI data analysis projects



	Preprocessing	Processing	Postprocessing
Sensor data	25%	25%	50%
Omics data	30%	20%	50%
Image data	35%	30%	35%
Clinical data	80%	10%	20%

Typical applications of AI to solve medical image problems

Image classification

Diagnosis: Decide the patient condition

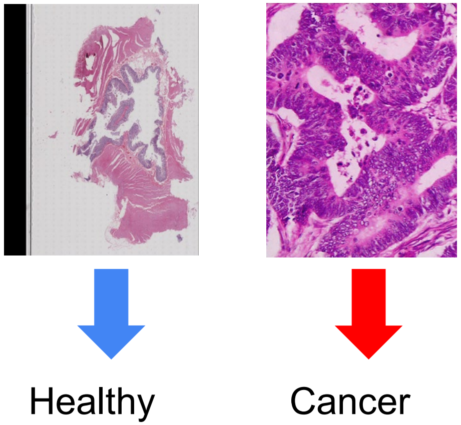
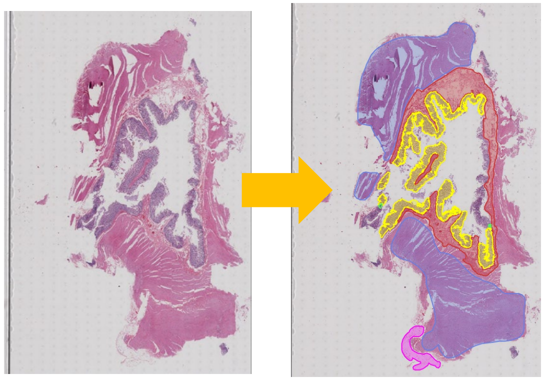


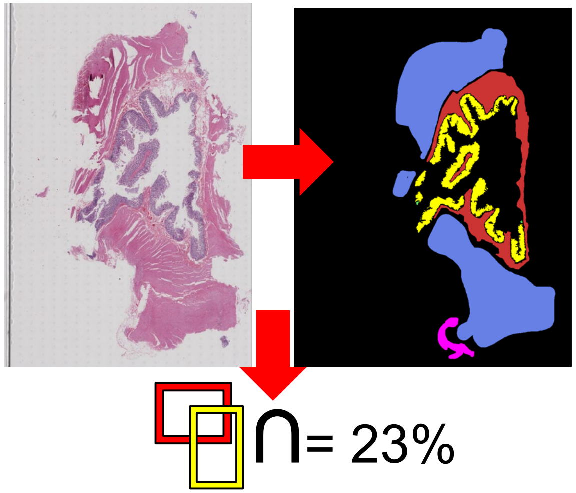
Image segmentation

Find regions in an image



Penetration calculation

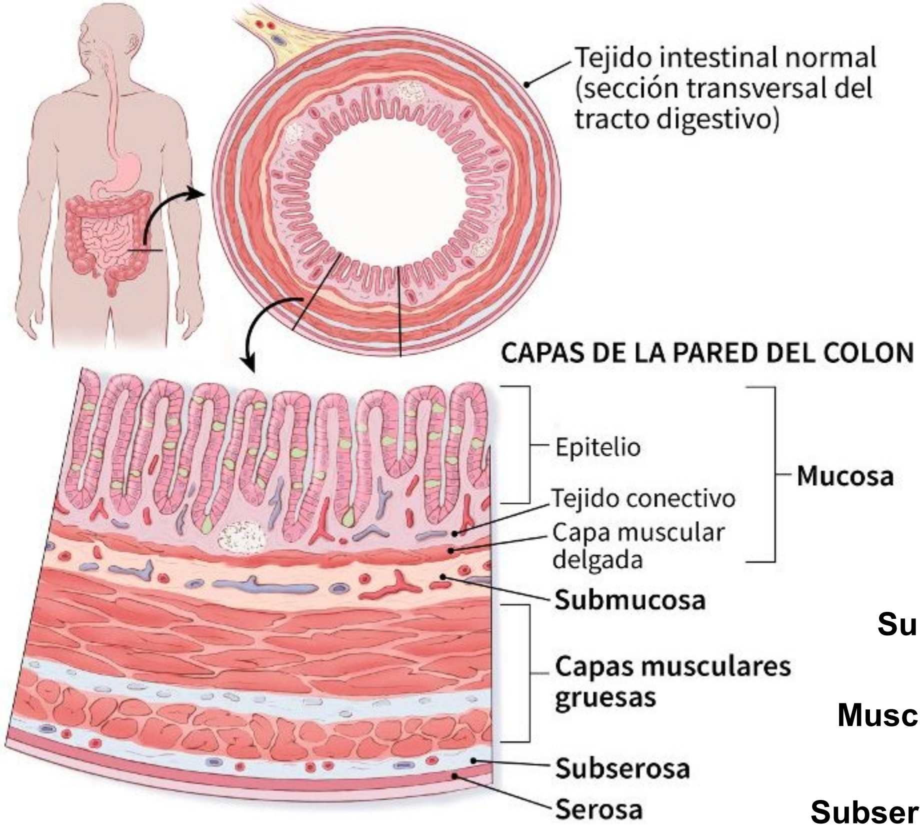
Calculate intersections of regions

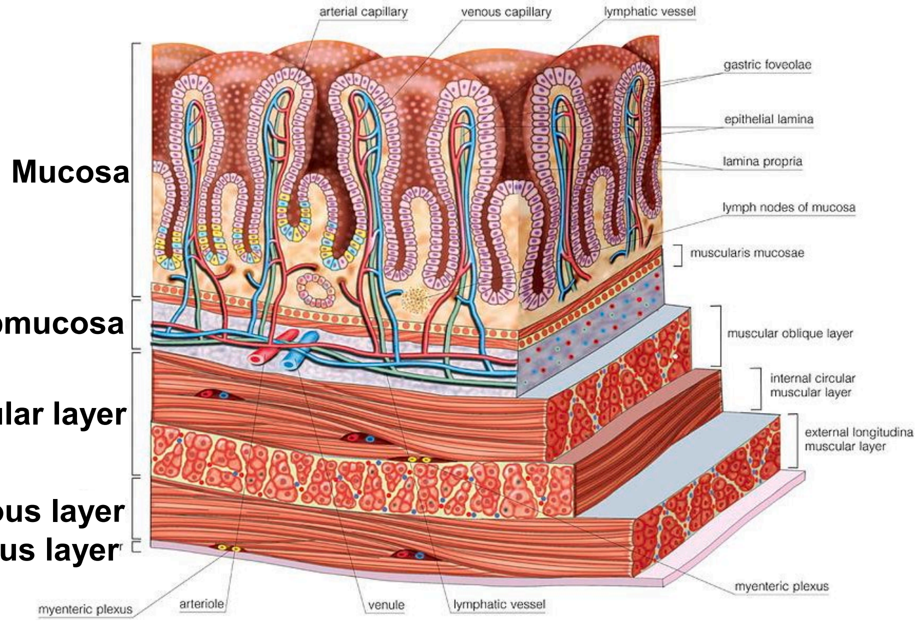


Challenges in the Design of Experiments (DoE) of medical images

- Medical information is not always electronic. Necessity to scan images.
- Image invalance: Much more healthy than disease images.
- Pathologists have **scarce time** to electronically record their decisions.
- Number of data adaptability: The DoF has to be robust to a reduction of the potential number of images of some categories.

Layers of the colon walls





Source: https://www.cancer.org/es/cancer/cancer-de-colon-o-recto/deteccion-diagnostico-clasificacion-por-etapas/clasificacion-de-la-etapa.html

Source: https://fineartamerica.com/featured/3-stomach-wall-asklepios-medical-atlas.html

Colon cancer stages

Stage 0: Cancer cells (CCs) are contained to the rectum's-colon's inner lining.
Abnormal cells are found in the innermost layer (**mucosa**), but have not become cancerous.

Stage 1: CCs are in deeper layers (colon-rectum wall), but they haven't spread beyond the wall.

- CCs are found in the innermost layer lining the colon-rectum. They have grown into the 2nd layer of tissue (**submucosa**).
- CCs may have also spread to a nearby muscle layer (**muscularis propria**) but hasn't reached nearby **lymph nodes** (**LNs**).

Stage 2: CCs have not spread to **LNs**, but have spread through and beyond the wall of the colon-rectum into nearby tissues, organs.
Stage 2A: CCs have spread through layers of colon-rectum wall & reached the **outermost layer**, but no farther.
Stage 2B: CCs have grown past outermost layer of colon-rectum wall but hasn't spread to nearby tissues or organs.
Stage 2C: CCs have spread past outermost layer of colon-rectum wall, grown into nearby tissues. Hasn't spread to **LNs** or distant organs.

Stage 3: CCs have spread to 1≥ nearby lymph nodes. Have not grown beyond **LNs**, colon-rectum wall to other parts of the body.
Stage 3A: CCs have spread through the 1st 2 inner layers of colon-rectum wall (**mucosa** & **submucosa**), may also reached the 3rd layer (**muscularis propria**).

- It has also reached 1-3 nearby **LNs**.
- Or has spread through the first two layers of the colon-rectum wall & has reached 4-6 nearby **LNs** .

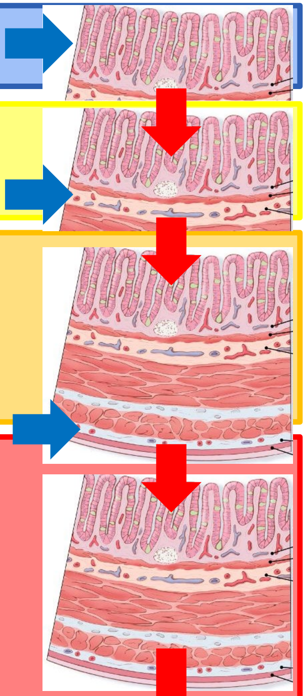
Stage 3B: CCs have reached the outermost layer (**serosa**) of the colon-rectum wall. It may have spread through the tissue that lines the abdominal organs (**visceral peritoneum**) but has not yet reached nearby organs.

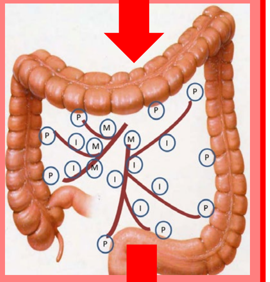
- CCs are found in 1-3 nearby **LNs**.
- Or has grown into the muscle layer or the outermost layer of the colon-rectum wall & has reached 4-6 nearby **LNs**.
- Or has grown through the 1st 2 layers of the colon-rectum wall & may have reached the muscle layer. CCs ares found in 7≥ nearby **LNs**.

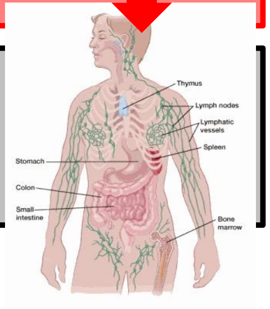
Stage 3C: CCs have grown past the colon-rectum wall & has spread to the tissue that lines abdominal organs. Has not spread to nearby organs..

- CCs are found in 4-6 nearby **LNs**.
- Or has grown past the colon-rectum wall or spread through the tissue that lines abdominal organs. It's found in 7≥ nearby **LNs**.
- Or has spread past the wall of the colon-rectum & has grown into nearby organs. CCs are found in 1≥ nearby **LNs**.

Stage 4: CCs have spread beyond the colon-rectum to distant areas of the body, including tissues and/or organs.
Stage 4A: CCs have reached one area or organ that isn't near the colon or rectum (liver, lung, ovary, faraway **LNs**.
Stage 4B: CCs have reached more than one area or organ that isn't near the colon-rectum.
Stage 4C: CCs have spread to distant parts of the tissue that lines the abdominal wall & may have reached other areas or organs.







Source: https://www.cancercenter.com/cancer-types/colorectal-cancer/stages

Images of a possible history of a colon cancer

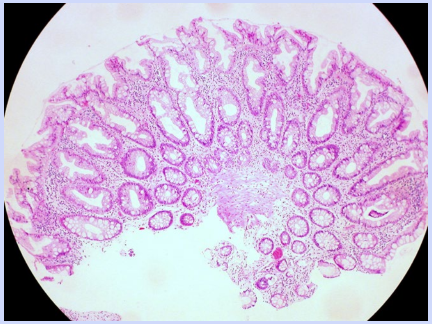
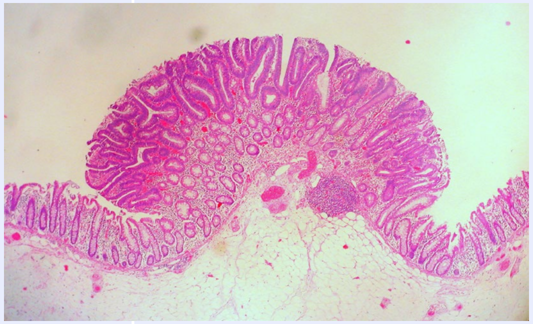
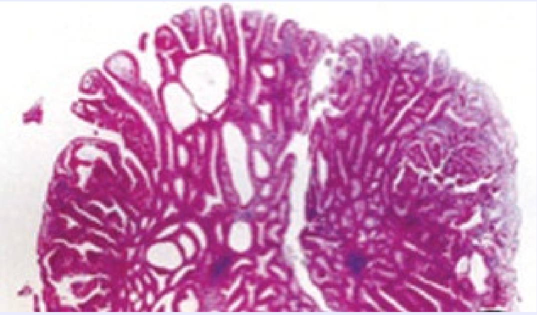
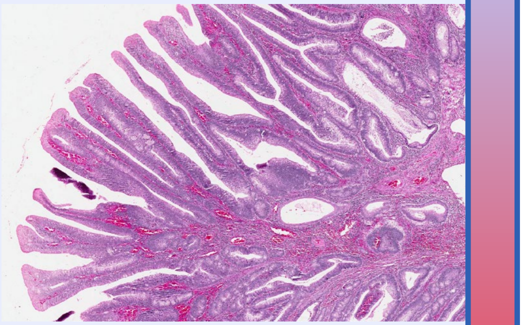
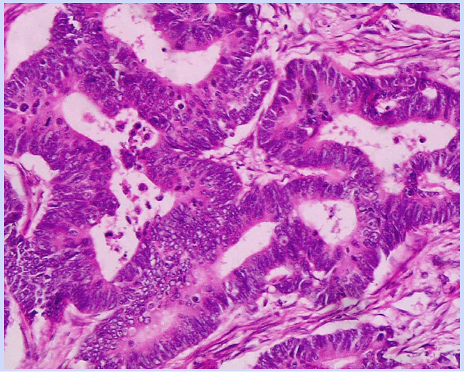
Type	Subtype (risk of containing malignant cell)
Hyperplastic polyp	(0%) 
Adenoma	Tubular adenoma(2%) Tubulovillous adenoma(20%-25%) Villous adenoma(15%-40%   
Colorectal adenocarcinoma	(100%) 

Image Codification

Input

Segmentation

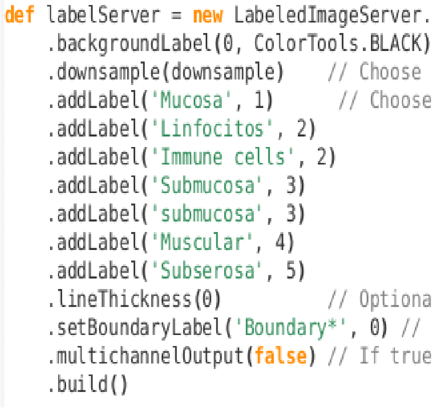
3	3	3	3	3	3	3	3	3	3	3	3	5	5	5	5	5	5
3	3	3	3	3	3	3	3	3	3	3	3	5	5	5	5	5	5
3	3	3	3	3	3	1	1	3	3	3	3	5	5	5	5	5	5
3	3	3	3	3	1	1	1	1	3	3	3	5	5	5	5	5	5
3	3	3	3	3	3	1	1	3	3	3	5	5	5	5	5	5	5
5	5	3	3	3	3	1	1	3	3	5	5	5	5	5	5	5	5
4	4	3	4	1	1	1	1	1	1	4	4	4	5	5	5	5	5
4	4	3	4	1	1	1	1	1	1	4	4	4	4	4	5	5	5
4	4	4	1	1	1	1	1	1	1	1	4	4	4	4	4	4	4
3	3	3	1	1	1	1	1	1	1	1	4	4	4	4	4	4	4
3	3	3	1	2	2	1	1	1	1	1	4	4	4	4	4	4	4
3	3	3	1	2	2	1	1	1	1	1	4	4	4	4	4	4	4

1: Person
2: Purse
3: Plants/Grass
4: Sidewalk
5:Building/Structures

Semantic Input

- The image must be codified with a **number code**: Each pixel has the value of the **class** it belongs to.
- This create a **mask of integers**.
- This mask is used as **supevision information during the training phase**,
- and as **output prediction during the validation phase**.

Image Exporting Script: Define the dictionary of classes



```
def labelServer = new LabeledImageServer.
    .backgroundLabel(0, ColorTools.BLACK)
    .downsample(downsample)    // Choose
    .addLabel('Mucosa', 1)      // Choose
    .addLabel('Linfocitos', 2)
    .addLabel('Immune cells', 2)
    .addLabel('Submucosa', 3)
    .addLabel('submucosa', 3)
    .addLabel('Muscular', 4)
    .addLabel('Subserosa', 5)
    .lineThickness(0)          // Optiona
    .setBoundaryLabel('Boundary*', 0) //
    .multichannelOutput(false) // If true
    .build()
```

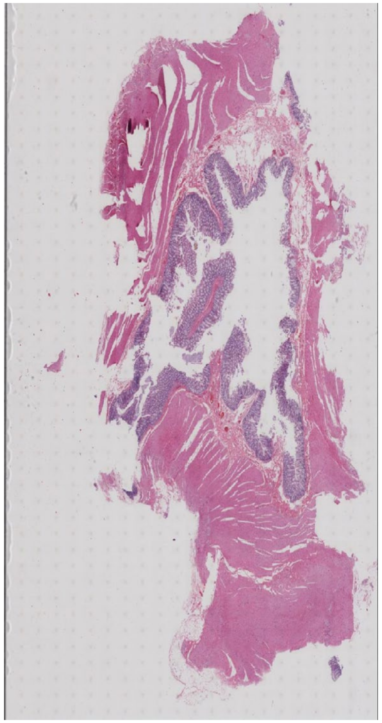
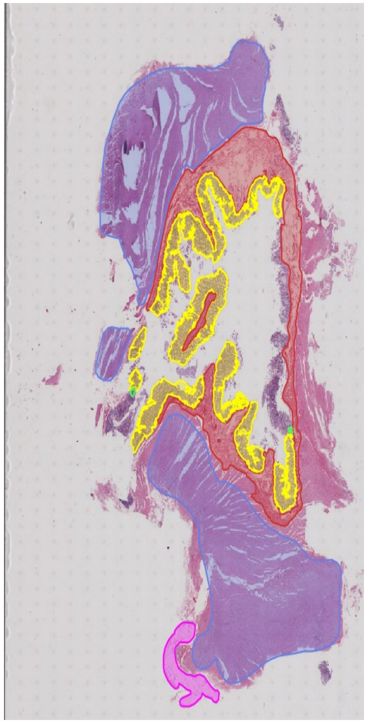
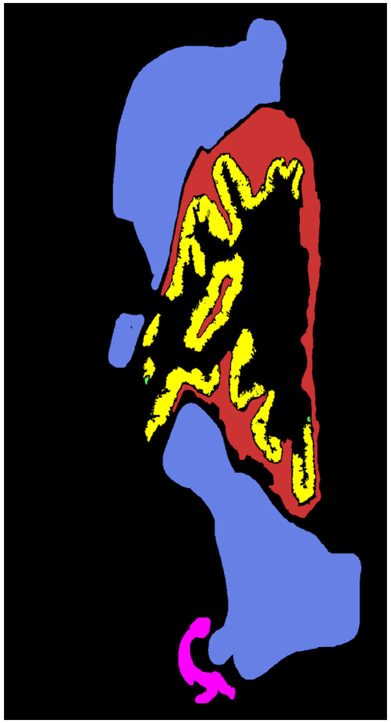


image: 10-6372 HEN



QuPath view



resulting mask

0: Background

1: Mucosa

2: Linfocites

3: Submucous

4: Muscular

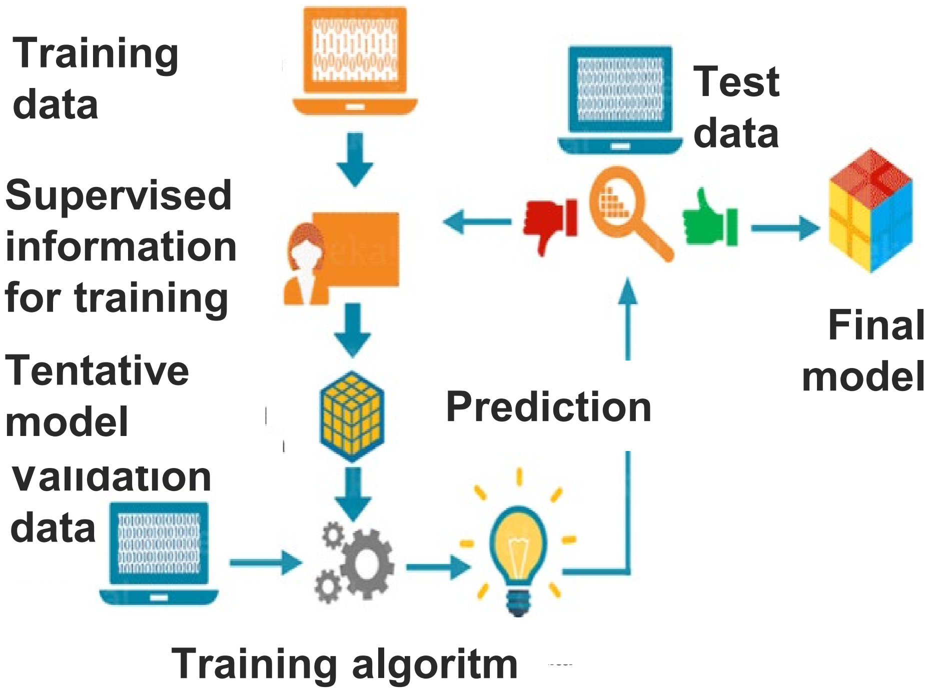
5: Subserous

...

❑Define the dictionaries in coordination with the pathologists, with such information will be create the supervsion information of the network.

- **More than one tissue type can be labelled with the same number** but **all the possibilities have to be defined in advance**.
- If there is any **overlap only the last exported** tissue will appear in th mask.

Training Process



Training data

Test data

Supervised information for training

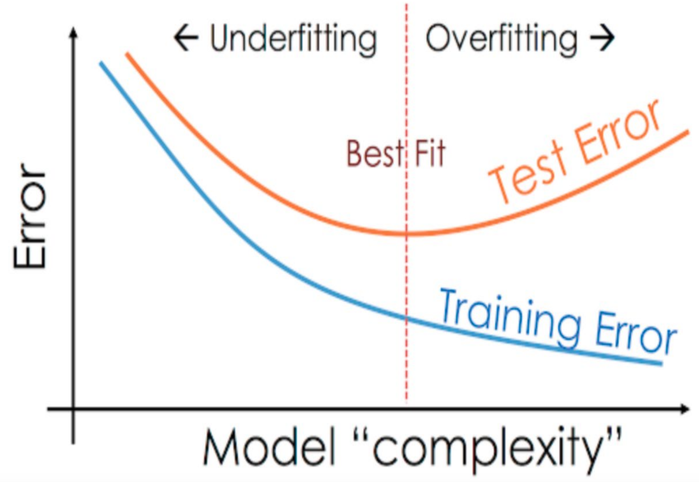
Final model

Tentative model

Prediction

Validation data

Training algoritm



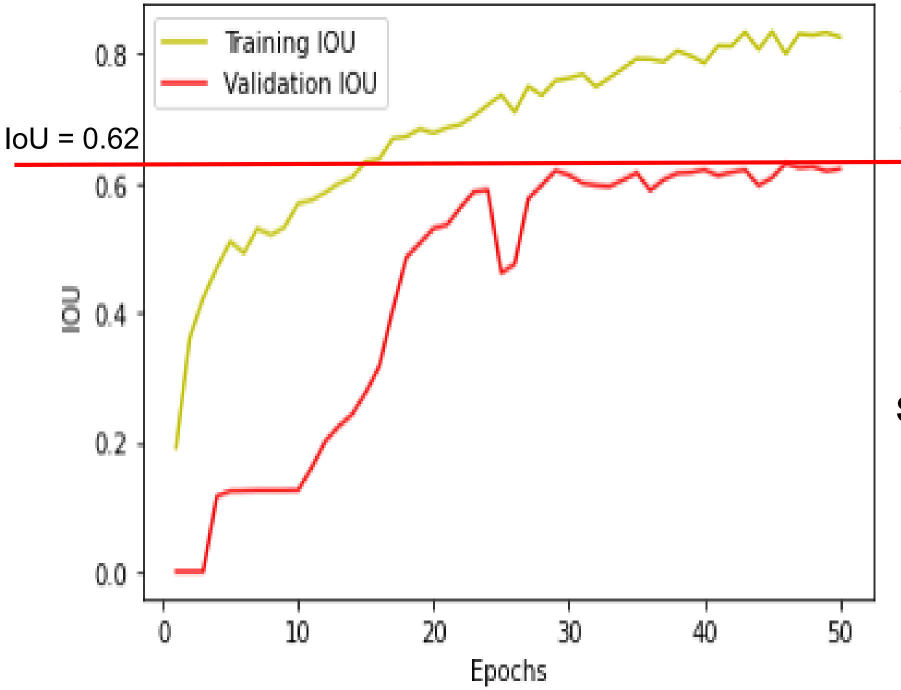
To have the opportunity to learn hidden information it could be interesting to increase the image resolution.

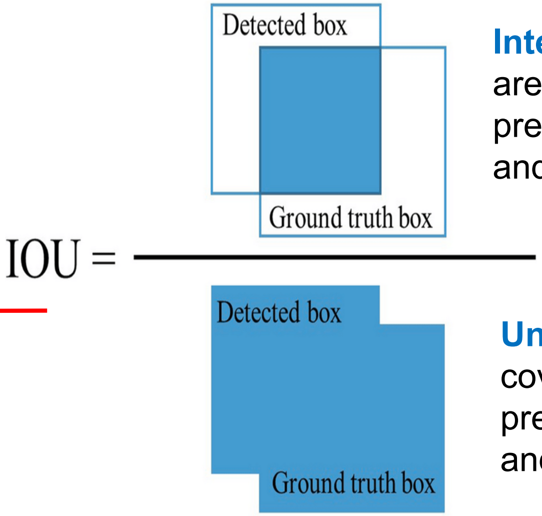
- ❑ **Learning drawbacks**: However, this could increase the **overfitting** if we do not include additional images for training.
- ❑ **Hardware drawbacks**: Increase computational demand (**GPU** & **storage**) in proportion to the square of the resolution.

To split the data en **training**, **validation** and **test** sets reduce the available data for training. Posible solution: **Jack-knife**: Trainf with *n*-1 images *n* models.

Intersection over Union (IoU) segmentation metric

Training and validation IoU (Higher is better)

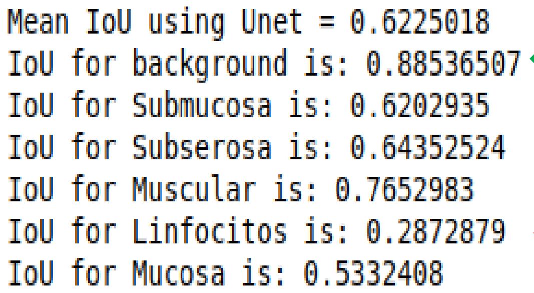




Intersection of the area between the predicted tissue and the real tissue

Union of the area covered by the predicted tissue and the real tissue

Some initial results



```
Mean IoU using Unet = 0.6225018
IoU for background is: 0.88536507
IoU for Submucosa is: 0.6202935
IoU for Subserosa is: 0.64352524
IoU for Muscular is: 0.7652983
IoU for Linfocitos is: 0.2872879
IoU for Mucosa is: 0.5332408
```

OK

- The **overfitting** happens if the performance index, Intersection over Union (IoU) **decrease with the number of training epochs**.
- By the moment we do not observe overfitting, however it is **very important to include additional images for training**.

Initial results on Healthy tissue detection

Resolution:

128x128 pixels

- 🟪 Background
- 🟦 Subserous
- 🟦 Muscular
- 🟨 Mucous
- 🟦 Submucous
- 🟩 Linfocites

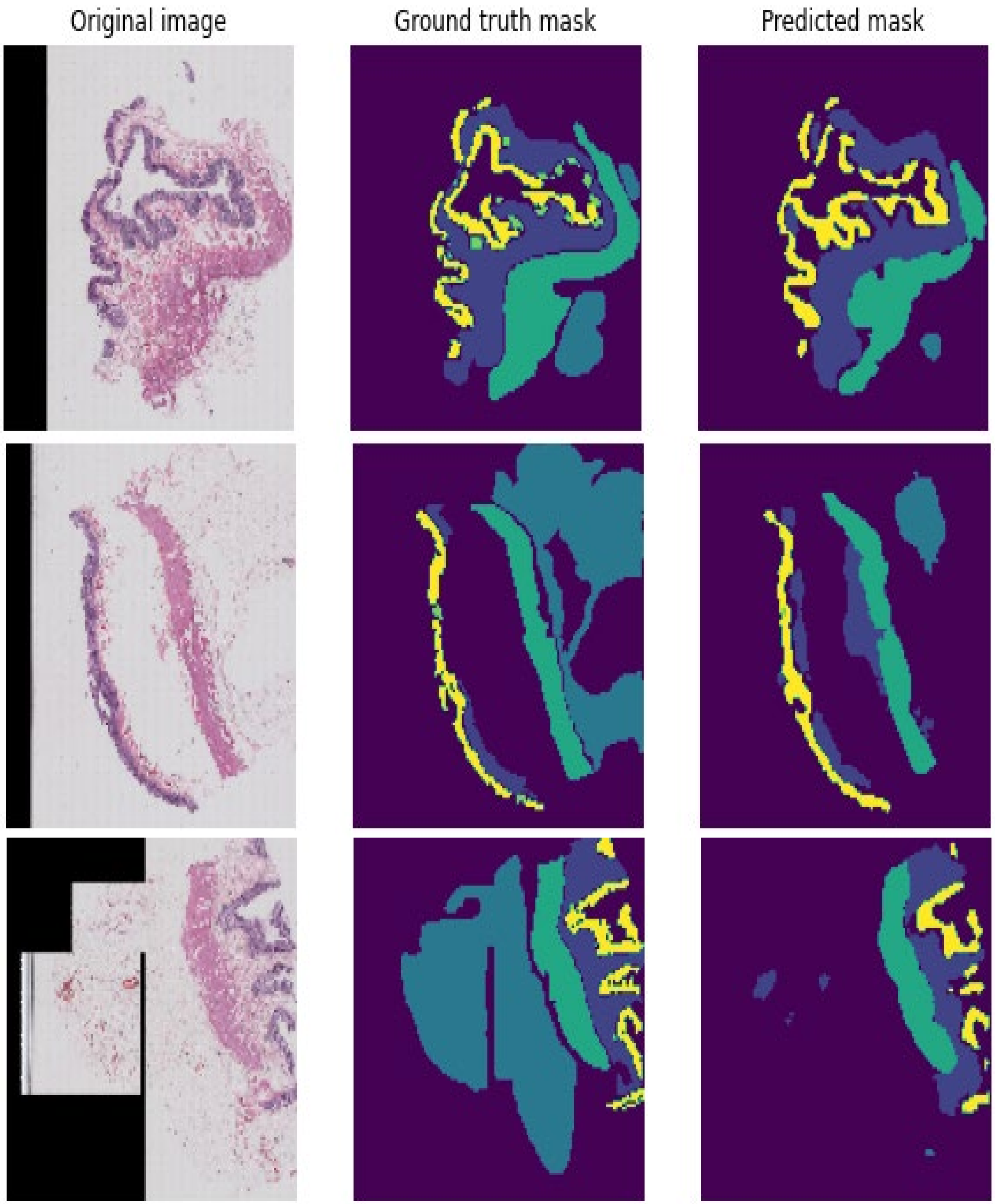


image: 10-2266 HEN

image: 10-2062 HEN

image: 10-1960 HEN

Results with higher resolution

Resolution:

512x512 pixels

- 🟪 Background
- 🟦 Subserous
- ⬛ Muscular
- 🟨 Mucous
- 🟦 Submucous
- 🟩 Linfocites

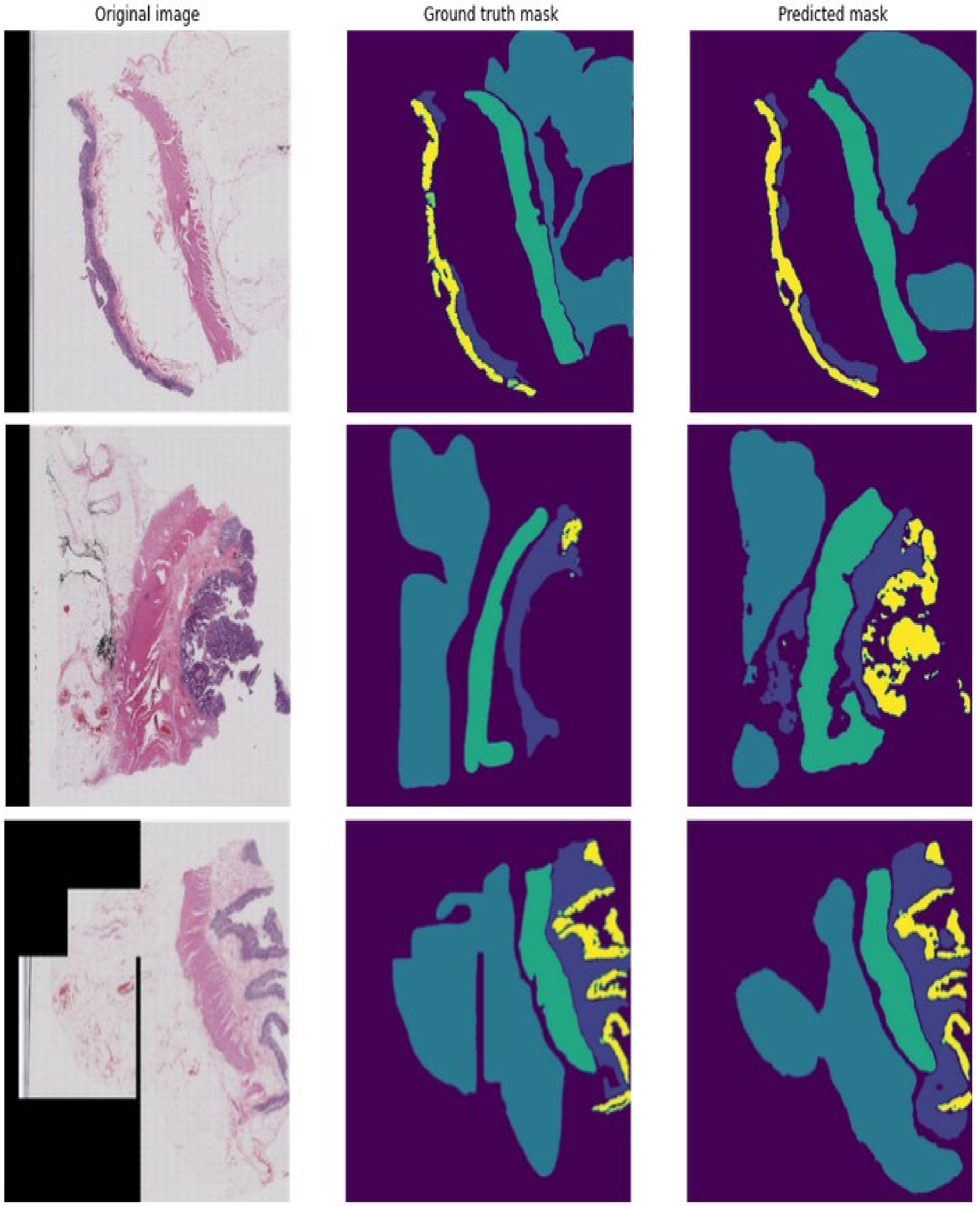


image: 10-2062 HEN

image: 10-7346 HEN

image: 10-1960 HEN

Resolution Comparison

128 x 128 pixels

- 🟪 **Background**
- 🟦 **Subserous**
- ⬛ **Muscular**
- 🟨 **Mucous**
- 🟦 **Submucous**
- 🟩 **Linfocites**

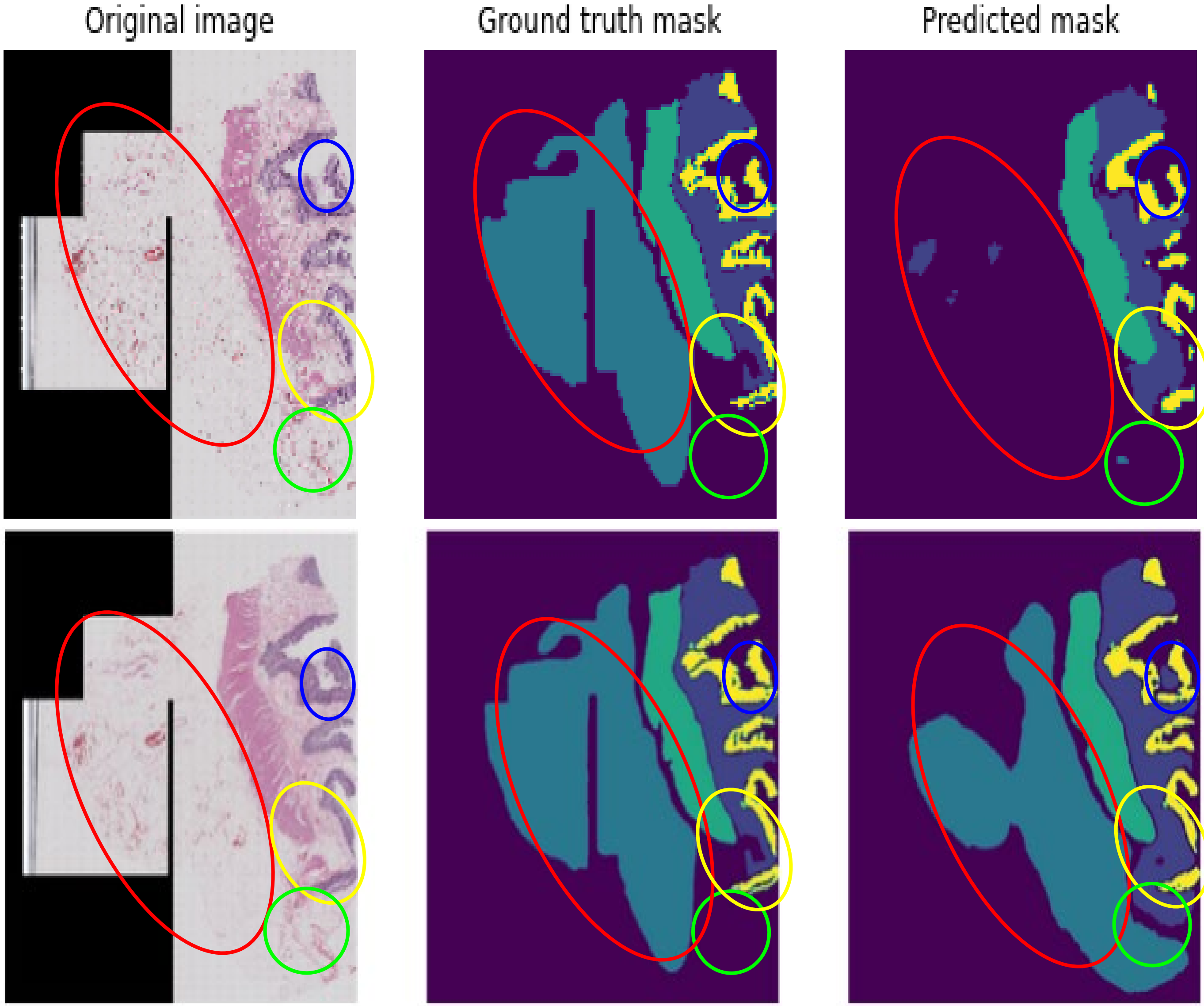


image: 10-1960 HEN

512 x 512 pixels

image: 10-1960 HEN

❑ Higher resolution discover **Muscular** tissues

New no marked detections

Resolution:
512x512 pixels

🟪 **Background**

🟦 **Subserous**

⬛ **Muscular**

🟨 **Mucous**

🟦 **Submucous**

🟩 **Linfocites**

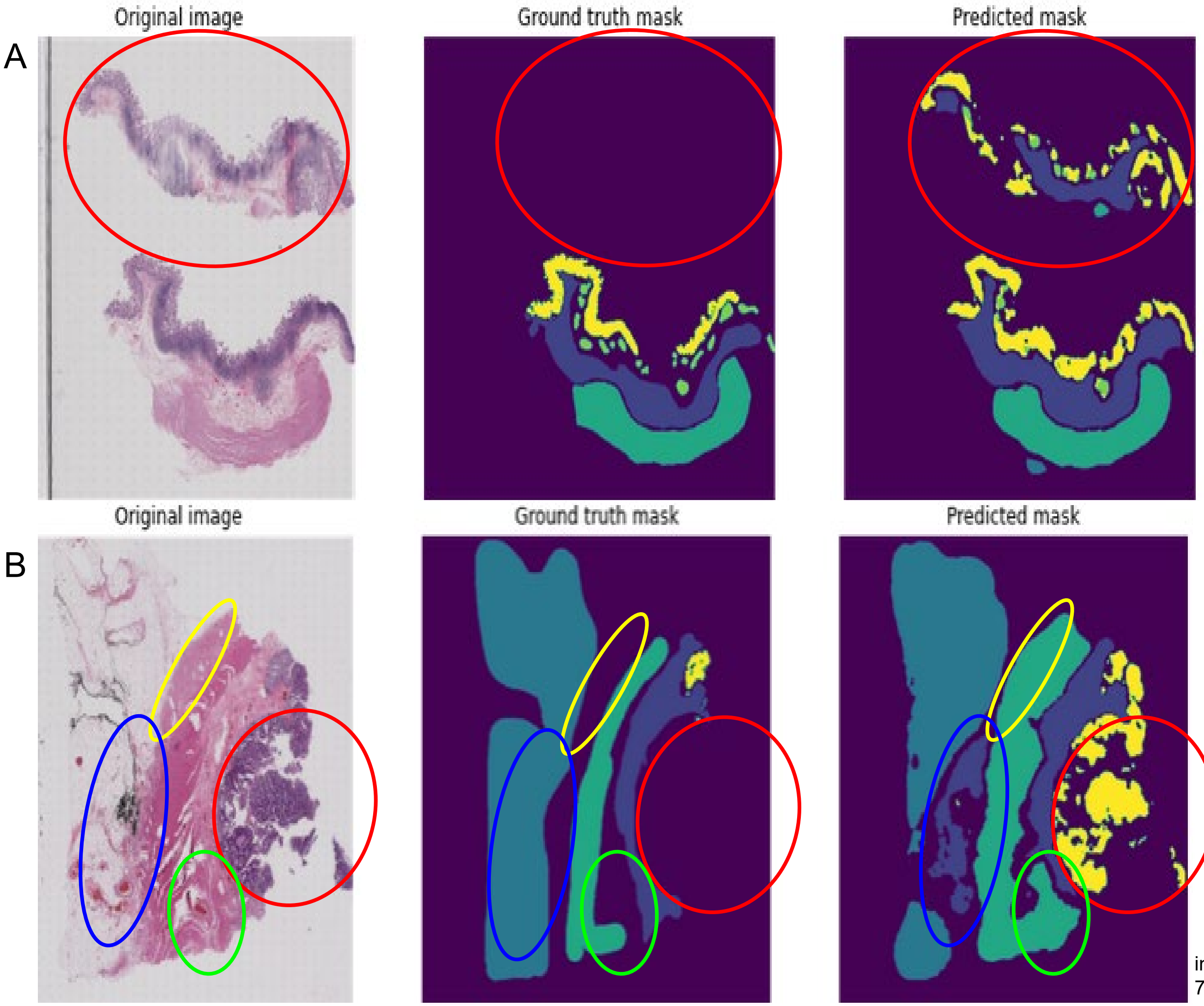


image: 10-3287 HEN

image: 10-7346 HEN

A. Discovering **whole regions** not marked by the pathologists.
B. Discovering of **Mucous**.

Next steps that we are implementing now

- **Train deep layers** with biomedical images to fine tune the **U-Net** segmentation architecture using the **ResNet34** backbone pre trained for the **ImageNet** dataset.

- **Quantify the percentage of predicted tissue** in each imagen.

- Define the **degree of confidence** on each predicted class, now we are using the Shannon's entropy.

- Define and represent the **performance indexes** (% of confidence in each class)
  - In order to compare the performance between different implementations, topologies, parameters etc.

- **Open the CNN black box**: Show neurons activity involved in each prediction using tools such as **Microscope**.

- **Software distribution** (predictor, not learner) by a Graphic User interface (GUI). Several ways of distribution:
  - Python
  - Java
  - Python or Java inside a container (Docker)
  - Inside server (in case of slow predictor)

Even the predictor is very computational demanding for the typical computer of a pathologist ⚠

Conclusions

- The non annotated tissue parts could be classified as “**non classified**” class.
  - Anyway, could be interesting to evaluate how the system is performing at predicting this unclassified tissue, since the main property of neural networks is being able to perceive patterns people can't.

- Since image full annotation is a very time consuming task for the pathologists (∼30 minutes/image), in carcer images could **only be annotated the cancer tissue**.
  - However, having a part of the cancer images fully annotated (all tissue types) could help the system to “understand the context” and therefore improve the performance.

- **Grade the tissues** that are most important in cancer invasion, so we focus more on the prediction performance of the more **critical tissues**.

- The main factor to improve the system performance is to have the **maximum amount of example images** for training from both healthy tissue and cancer.

Acknowledgments

biodonostia
osasun ikerketa institutua
instituto de investigación sanitaria

Computational Biology & Systems Biomedicine

- ✓ **Julen Bohoyo Bengoetxea**

- ✓ Daniela Gerovska
- ✓ Alex Martinez Ascensión
- ✓ Mikel Arrospide Elgarresta
- ✓ Javier Cabau Laporta

bioaraba
osasun ikerketa institutua
instituto de investigación sanitaria

Keralty

- ✓ **Jose J. Rodriguez Anda**

Always looking for motivated programmers

Marcos J. Araúzo-Bravo (mararabra@yahoo.co.uk)

HORIZON 2020

COST

ikerbasque
Basque Foundation for Science

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

ERACO SYSMED

ORAIN GIPUZKOA

EUSKO JAURLARITZA
GOBIERNO VASCO

HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

HUMAN CELL ATLAS

1957